Aide à la décision dans les filières agroalimentaires

Rallou Thomopoulos

Habilitation à Diriger des Recherches
Université Montpellier II, école doctorale I2S, spécialité informatique

Aide à la décision
dans les filières agroalimentaires

Decision support in agrifood chains

Rallou THOMOPOULOS

INRA, UMR IATE
Equipe-projet INRIA GraphIK
Montpellier, France

Soutenue le 5 décembre 2013

JURY

Christine FROIDEVAUX	Professeur, Université Paris Sud, Orsay	Rapporteur
Bernard MOULIN	Professeur, Université Laval, Québec, Canada	Rapporteur
Henri PRADE	Directeur de Recherche CNRS, IRIT, Toulouse	Rapporteur
Joël ABECASSIS	Ingénieur de Recherche (HDR), INRA, Montpellier	Examinateur
Fabien GANDON	Chargé de Recherche (HDR), INRIA, Sophia-Antipolis	Examinateur
Marie-Laure MUGNIER	Professeur, Université Montpellier II	Présidente du jury

Résumé

Dans les sciences expérimentales telles que les sciences de l'aliment, les données jouent un rôle essentiel, puisque les théories du domaine sont fondées sur les données expérimentales, leur exploitation et leur analyse. Cependant, l'état de l'art montre que les données expérimentales disponibles sont souvent partielles, éparpillées sur des supports variés, ou sans modèle mathématique sous-jacent établi. Une autre source d'information est également disponible : les connaissances expertes, toutefois pas toujours formalisées sur des supports écrits. Les connaissances expertes peuvent exprimer des points de vue différents, potentiellement conflictuels s'ils visent des objectifs divergents. Un défi majeur est donc d'intégrer ces données et ces connaissances et de développer des méthodes permettant de les utiliser pour l'aide à la décision. Ce mémoire présente un ensemble de stratégies et méthodes complémentaires définies et développées pour, ensemble, traiter cette problématique. Il aborde trois thèmes de recherche : l'*intégration de formalismes hétérogènes*, les *méthodes prédictives* et l'*argumentation pour l'aide à la décision*.

Abstract

In experimental sciences such as food science, data play an essential role, since domain theories are based on experimental data, their exploitation and their analysis. However, the state of the art shows that available experimental data are often partial, scattered on various supports, or without an established underlying mathematical model. Another information source is also available : expert knowledge, however not always formalized on written supports. Expert knowledge may express different viewpoints, possibly conflictual since they pursue divergent objectives. A main challenge is thus to integrate these data and knowledge and to develop ways of supporting decision from them. This research report presents a set of complementary strategies and methods defined and developed in order to, together, face this issue. It addresses three research topics : *integration of heterogeneous formalisms*, *predictive methods*, and *argumentation for decision support*.

Remerciements

Je souhaite en premier lieu remercier les membres du jury pour l'ouverture d'esprit dont ils ont fait preuve dans leur écoute, leurs questions et discussions sur ce thème pluridisciplinaire. Par leur contribution, cette soutenance d'HDR a été pour moi une journée extrêmement enrichissante, constructive et motivante. Je tiens à exprimer le fait que chacun d'entre eux, par différents contextes dans lesquels je les ai rencontrés, a contribué à construire mon parcours. Je remercie particulièrement les rapporteurs pour leur lecture approfondie du manuscrit et pour leurs rapports, mais aussi les autres membres du jury qui en ont fait une lecture attentive et minutieuse. Je les remercie également de s'être rendus disponibles malgré des emplois du temps extrêmement chargés.

J'aimerais remercier Bernard Cuq et Joël Abécassis pour le rôle qu'ils ont joué lors de mon arrivée à l'UMR IATE. Bien que de disciplines différentes et malgré le temps nécessaire pour comprendre les compétences de chacun, leur implication a permis à mon projet initial de prendre un bon départ, de se poursuivre et de se développer par la suite. Je tiens également à remercier Stéphane Guilbert, directeur de l'UMR IATE lors de mon recrutement, qui, conscient de la difficulté d'initier un projet de recherche dans une discipline nouvelle pour l'unité, m'a beaucoup apporté par ses conseils d'ordre stratégique et sa confiance. Merci à tous trois de m'avoir accompagnée lors de mon arrivée.

Je remercie très sincèrement les personnes avec qui j'ai eu le plaisir de travailler : l'axe 5, l'équipe-projet INRIA GraphIK, les membres de l'UMR MISTEA ainsi que les collègues du LRI et de l'IRIT – en particulier Fatiha Saïs et Leïla Amgoud –, les étudiants qui ont contribué à ces travaux, ainsi que l'ensemble des membres de l'UMR IATE pour leur accueil et leur sympathie. Un grand merci à Madalina Croitoru, Jérôme Fortin, Brigitte Charnomordic, Nadine Hilgert et Patrice Buche qui m'ont aidée à améliorer mon exposé oral. Un mot tout particulier à Brigitte Charnomordic, avec qui nous avons eu de nombreuses occasions de collaboration, que je remercie pour sa bonne humeur et sa très grande humanité.

Enfin, je glisse un petit merci à mon cocon familial, qui ne me juge pas quelque choix que je fasse.

Table des matières

Table des figures **10**

1 Introduction **11**

1.1 Contexte 11

1.2 Thèmes de recherche 14

1.3 Organisation du mémoire 18

2 Intégration de formalismes hétérogènes **19**

2.1 Problématique 20

2.2 Notions préliminaires 22

2.2.1 Les graphes conceptuels simples 22

2.2.2 Règles de graphes conceptuels 25

2.3 Génération d'une ontologie 26

2.3.1 Travaux proches 27

2.3.2 Identification de types de concepts de haut niveau . . . 29

2.3.3 Hiérarchisation des types de concepts 31

2.3.4 Proposition de types de concepts complémentaires . . . 33

2.4 Evaluation de la validité des dires d'experts 34

2.4.1 Problématiques proches 34

2.4.2 Calcul du taux de validité 35

2.4.3 Notions de patron de règle, d'instance de règle et propriétés associées 35

2.4.4 Déroulement de la validation d'une instance de règle . 40

2.5 Application 42
2.5.1 Environnement de travail 42
2.5.2 Description des données expérimentales 43
2.5.3 Description des connaissances expertes 43
2.5.4 Validation des connaissances expertes 44
2.6 Conclusion du chapitre 45

3 Méthodes prédictives 47
3.1 Problématique 48
3.2 Littérature pertinente 51
3.2.1 Utilisation d'ontologies pour guider l'apprentissage . . 51
3.2.2 Utilisation de l'analyse subjective pour la sélection de règles ou de données 52
3.2.3 Les arbres de décision comme modèles interprétables . 53
3.3 Définition de l'ontologie en lien avec les données 54
3.3.1 Domaine de définition des concepts 56
3.3.2 Relation entre concepts et variables 56
3.3.3 L'ensemble des relations 56
3.4 Traitement des données utilisant l'ontologie 59
3.4.1 Remplacement d'une variable par de nouvelles variables 61
3.4.2 Regroupement de modalités d'une variable sur la base de propriétés communes 62
3.4.3 Fusion de variables pour créer une nouvelle variable . . 62
3.5 Approche interactive : principes et évaluation 63
3.5.1 Principes 63
3.5.2 Evaluation 64
3.6 Etude de cas : application à la prédiction de la qualité alimentaire 64
3.6.1 Contexte et description de l'étude de cas 65
3.6.2 Application de l'approche à l'étude de cas 65
3.7 Conclusion du chapitre 70

4 Argumentation pour l'aide à la décision 71

4.1 Problématique 71
4.2 Méthodologie 74
4.2.1 Identification et analyse des sources d'information . . . 74
4.2.2 Modélisation des informations disponibles en arguments structurés 75
4.2.3 Modèles d'argumentation existants 76
4.3 Résultats 80
4.3.1 Schéma global 80
4.3.2 Arguments 81
4.3.3 Le modèle proposé 84
4.3.4 Actions recommendées pour d'autres préoccupations et d'autres audiences 87
4.4 Conclusion du chapitre 88

5 Une méthode d'ingénierie inverse pour le pilotage de filière 91
5.1 Problématique 91
5.2 Les éléments du formalisme 93
5.2.1 Pourquoi un langage logique ? 93
5.2.2 Définitions en logique du premier ordre 94
5.2.3 Conséquence logique, substitution et homomorphisme . 95
5.2.4 Règles et dérivation 95
5.2.5 Expression de l'inconsistance 97
5.2.6 Base de connaissances consistante 98
5.2.7 Réponse à une requête : chaînage avant et chaînage arrière 99
5.3 Modéliser le problème 99
5.3.1 Présentation du cas d'étude 99
5.3.2 Exprimer les caractéristiques-cibles suivant différents points de vue 100
5.3.3 Formalisation des buts 101
5.3.4 Traduire l'ingénierie inverse 103
5.3.5 Formalisation du processus d'ingénierie inverse 103

5.3.6 Exemple illustratif 104
5.4 Aide à la décision 105
5.4.1 Calcul des arguments et des extensions 106
5.4.2 Choix des points de vue à retenir 108
5.5 Synthèse et discussion 109
5.5.1 Schéma global de la démarche 109
5.5.2 Autres approches et positionnement 111
5.6 Conclusion du chapitre 112

6 Perspectives **115**
6.1 Décision argumentée : vers une approche graphique et collaborative 115
6.2 Argumentation et analyse multidimensionnelle 118
6.3 Qualité des données et fusion de données redondantes 121

Bibliographie **123**

Table des figures

2.1 Un graphe conceptuel simple G 23
2.2 Une règle de graphe conceptuel simple R 25
2.3 Exemple de hiérarchisation des types de concepts 32
2.4 Exemple de règle experte de même forme que celle de la figure 2.2 36
2.5 Exemple de patron de règle 36
2.6 Une partie du vocabulaire utilisé pour exprimer les connaissances expertes 44
2.7 Evaluation de la validité d'une règle experte 45
2.8 Affichage des exceptions d'une règle experte 46

3.1 Schéma du processus de construction de modèle 50
3.2 Un extrait de l'ontologie utilisée pour les procédés alimentaires 55
3.3 Quelques variables et parties de l'ontologie associées, où $A \rightarrow B$ signifie que A est une *sorte de* B 60
3.4 Arbres de décision sur les données brutes 66
3.5 Arbres de décision sur les données avec les propriétés des vitamines 67
3.6 Arbre de décision avec le type de cuisson et les propriétés de l'eau 68
3.7 Arbre de décision à l'état final 69

4.1 Graphe d'attaque associé 78
4.2 Graphe avec attaques non-symétriques 80
4.3 Graphe d'attaque pour la préoccupation nutritionnelle 86

4.4 Graphes d'attaque propres à chaque audience 87

5.1 Buts nutritionnels 101
5.2 Buts organoleptiques 102
5.3 Moyens d'atteindre les buts nutritionnels 103
5.4 Schéma global de la démarche 110

6.1 Première partition 120
6.2 Seconde partition 120

Chapitre 1

Introduction

Dans cette introduction je m'attacherai à expliquer les choix de recherche que j'ai faits, leur dynamique et leur cohérence en fonction du contexte et des problématiques rencontrées depuis mon recrutement à l'INRA fin 2004.

1.1 Contexte

J'ai été recrutée au sein de l'UMR Ingénierie des Agropolymères et Technologies Emergentes (IATE) pour initier un nouvel axe de recherche : la représentation et l'intégration des connaissances. J'étais donc au départ seule chercheur en informatique dans l'unité.

Point critique lors de mon recrutement : le problème d'ingénierie inverse dans une filière

Lors des premières années de mon recrutement, l'enjeu des travaux menés par l'UMR IATE est une meilleure maîtrise des procédés de transformation, dans le but de pouvoir garantir la qualité et la sécurité des produits alimentaires. En particulier, la filière blé dur apparaît relativement bien connue et maîtrisée, si bien qu'une approche par ingénierie inverse (c'est-à-dire consistant à moduler le procédé de fabrication en fonction de propriétés-cibles visées pour le produit final) semble envisageable. En effet, les travaux menés dans l'UMR ont conduit à l'accumulation de données issues de sources et de disciplines variées (aspects technologiques, nutritionnels, organoleptiques,

etc.) ; c'est la capacité d'intégration de ces connaissances qui apparaît comme un objectif prioritaire pour permettre l'aide à la décision. Mon projet initial porte plus spécifiquement sur les informations concernant l'influence des matières premières et des procédés de transformation sur la qualité des produits à base de céréales et notamment de blé dur. On voit que les mots-clés "intégration" et "décision", qui définissent les thématiques principales de mes travaux actuels, sont déjà présents dans ce projet initial.

Ce que montre l'analyse de l'existant : une problématique de "données pauvres"

L'analyse de l'existant montre alors que les données expérimentales disponibles sont souvent partielles car dédiées à un problème très spécifique, éparpillées au travers de supports divers (articles, cours, rapports, brevets, ...), souvent sans modèle mathématique connu. Une autre source d'information est également disponible : les connaissances expertes, toutefois pas toujours formalisées sur des supports écrits. Cet état de fait explique une des premières tâches entreprises au cours de mon projet : contribuer à l'informatisation des données et connaissances disponibles via des projets permettant le développement de bases de données et de connaissances (KB-filière, Grain Virtuel, ...) et surtout une définition plus précise de mes priorités de recherche.

Conséquence : une orientation plus précise de mes priorités de recherche

Cette problématique de données pauvres m'a amenée à m'intéresser prioritairement à deux thèmes de recherche :

- dans un souci de cohérence des informations disponibles : l'*intégration de formalismes hétérogènes* ;
- dans un objectif de *prédiction* : l'étude de méthodes d'apprentissage pertinentes en situation de données pauvres, en particulier des approches de type arbres de décision et raisonnement à partir de cas.

La problématique d'intégration des connaissances s'est ensuite posée de façon plus large comme une question d'aide à la décision conciliant des points de vue contradictoires à l'échelle des filières, ouvrant un troisième thème de recherche :

- l'*argumentation* pour l'*aide à la décision*.

Ces trois thèmes sont développés dans la partie 1.2 ci-dessous.

Collectif de recherche

Une grande partie de mes travaux sont issus de la collaboration régulière (sous la forme de réunions hebdomadaires) avec l'équipe Représentation des Connaissances et Raisonnement du LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier), dont je suis membre associé depuis janvier 2006. Cette association a abouti au montage de l'équipe-projet INRIA-CNRS-UM2-INRA "GraphIK" (responsable : Marie-Laure Mugnier), localisée à Montpellier, officiellement créée le 1er janvier 2010, dont je suis membre permanent.

D'autre part, l'axe 5 (représentation des connaissances) de l'UMR IATE, dont j'étais seule chercheur en informatique lors de sa création fin 2004 (coïncidant avec mon recrutement), s'est progressivement développé par le recrutement d'un AI INRA (Luc Menut) en septembre 2007, par un chercheur CIRAD (Sébastien Destercke) en février 2009 puis par un IR INRA (Patrice Buche) et un MC UM2 (Jérôme Fortin) en septembre 2009, générant ainsi une dynamique d'équipe en représentation des connaissances au sein d'une unité applicative. La constitution de cette équipe a ainsi permis de rassembler des compétences scientifiques complémentaires sur les thèmes :

- de représentation et intégration de données et connaissances ;
- de raisonnement sur des données et connaissances numériques et symboliques.

Les méthodes développées, mises ensemble, permettent de bâtir une approche de plus en plus expressive, et crédible du point de vue applicatif, pour l'aide à la décision au sein des filières. Plus spécifiquement, ces compétences s'articulent de la façon suivante : l'acquisition semi-automatique de données à partir de sources hétérogènes (notamment tableaux de données de documents issus du web) permet de pallier, de façon complémentaire, le problème des données pauvres ; la prise en compte de l'incertitude concerne à la fois les modèles et les données (propagation d'incertitude dans les modèles, imperfection des données) ; le raisonnement non-monotone (e.g. logique des défauts) est mobilisable à la fois en prédiction dans une approche qualitative et comme technique de calcul d'extensions (ensemble d'arguments "cohérents") au sein des méthodes d'aide à la décision.

1.2 Thèmes de recherche

Premier thème : Intégration de formalismes hétérogènes

Les travaux entrepris traitent de la question de la coopération de connaissances hétérogènes pour la construction et la validation de l'expertise d'un domaine. Deux types de connaissances sont pris en compte :

- des dires d'experts, connaissances à caractère générique, découlant de l'expérience des spécialistes du domaine et décrivant les mécanismes communément admis régissant ce domaine. Ces connaissances sont représentées sous la forme de règles dans le modèle des graphes conceptuels, formalisme de représentation des connaissances fondé sur la logique, ayant notamment l'avantage, pour des utilisateurs non spécialistes, d'avoir une représentation graphique équivalente. Ce modèle comporte aussi une partie ontologique (vocabulaire hiérarchisé constituant le support du modèle) ;
- des données expérimentales, issues de la littérature internationale du domaine. Elles sont représentées dans le modèle relationnel. Ces données nombreuses décrivent avec précision et de façon chiffrée des expériences réalisées pour approfondir la connaissance du domaine et leurs résultats. Ces résultats peuvent ou non vérifier les connaissances apportées par les dires d'experts.

Ces travaux ont abouti, d'une part, à la *génération semi-automatique de connaissances ontologiques*. La méthode mise en place permet, à partir du modèle relationnel classique, d'identifier et de hiérarchiser (en termes de spécificité) des concepts de haut niveau (c'est-à-dire génériques) à partir des attributs, relations et données syntaxiques de la base de données. Cette extraction automatique, complétée par des concepts issus de l'application de techniques de fouille automatique de textes, a ensuite été complétée et validée par les experts de l'unité.

D'autre part, ces travaux ont permis d'étudier la communication entre informations hétérogènes : une information riche mais hétérogène (experte / expérimentale) n'est réellement utile que si des ponts sont constitués. Nous avons donc étudié la possibilité de mettre au point des méthodes de *validation et d'interrogation croisées*. Ces méthodes se proposent de résoudre deux problèmes :

- la validation de dires d'experts par les données : en s'appuyant sur les graphes conceptuels traduisant les connaissances expertes, un profil de règle est construit. Les prémisses de cette règle sont ensuite utilisées pour

interroger la base de données, et un taux de validité est calculé pour en déduire une éventuelle contradiction entre les données et la règle experte ;

- l'identification d'exceptions : les cas particuliers sont souvent importants, il s'agit de les identifier pour les prendre en compte. A partir de la méthode développée pour la validation, des cas contredisant une règle générique peuvent être identifiés, et des règles (dites de défaut) correspondant à ces cas peuvent ensuite être ajoutées au corpus de connaissances, en utilisant la connaissance hiérarchique fournie par l'ontologie. De telles règles sont la base de la *mise en oeuvre de logiques non-monotones*, dont l'intégration au cadre des graphes conceptuels est un sujet de recherche de l'équipe.

Ces travaux ont été réalisés en collaboration avec Jean-François Baget (INRIA GraphIK), Bernard Cuq (Supagro, UMR IATE) et Ollivier Haemmerlé (IRIT). L'intégration du raisonnement non monotone dans le modèle des graphes conceptuels a ensuite été développée avec Jérôme Fortin (UM2, IATE/GraphIK), Jean-François Baget (INRIA GraphIK) et Madalina Croitoru (UM2, GraphIK).

Les développements associés ont été réalisés dans le cadre des stages en informatique de Clément Molla (master 2), d'Amine Lakhoua (master 2) et de Samira Rezgane (IUP), que j'ai encadrés, ainsi que celui de Clotilde Raz (ingénieur en nutrition), encadrée par Bernard Cuq, pour la partie conception.

Un point de recherche subsidiaire concerne la fusion de références. Cette tâche vise à identifier puis fusionner des données qui réfèrent à la même entité du monde réel (doublons, données redondantes, ...) pour améliorer la qualité des données. L'approche proposée s'appuie sur les ensembles flous pour permettre une *représentation prenant en compte l'incertitude, ainsi qu'une interrogation flexible* – avec expression de préférences graduelles – des données fusionnées. Une implémentation utilisant les standards du W3C (représentation RDF et interrogation SPARQL) a été proposée. Ces travaux sont menés en collaboration avec Fatiha Saïs (post-doctorante dans l'équipe RCR du LIRMM, puis LRI/Université d'Orsay) et Sébastien Destercke (CIRAD UMR IATE, puis CNRS Heudiasyc). J'interviens également avec Fatiha Saïs sur la question de la fusion de références dans le projet ANR CONTINT "Qualinca" porté par Michel Leclère (UM2, GraphIK) et démarré en 2012.

Deuxième thème : Méthodes prédictives

Deux types de méthodes prédictives (raisonnement à partir de cas et arbres de décision) ont été retenus pour certaines de leurs propriétés : leur faible besoin en données, la prise en compte de données symboliques et numériques, la gestion des valeurs manquantes. Les deux types de méthodes prédictives abordés ont des avantages complémentaires : les *méthodes de raisonnement par cas* (ou par analogie) ont été retenues pour leur simplicité d'utilisation, leur faible besoin en données et leurs bons résultats ; les *méthodes inductives*, pour leur capacité à construire des modèles génériques et à mettre en évidence de nouvelles connaissances impliquant des relations complexes, dans lesquelles de nombreuses variables sont en interaction, difficilement décelables par l'expertise. Pour ces dernières, les *modèles graphiques* ont été privilégiés (les arbres de décision en l'occurrence) en raison de leur facilité d'interprétation.

L'originalité de l'approche fondée sur les arbres de décision est la définition d'une méthode de collaboration entre modèles prédictifs, connaissances ontologiques et connaissances expertes. Des expérimentations, effectuées sur le problème de l'évolution des différents types de vitamines au cours de la cuisson des pâtes alimentaires, ont permis de confirmer l'intérêt de ces méthodes. La méthode a été développée en collaboration avec Brigitte Charnomordic (INRA UMR MISTEA) et Sébastien Destercke (CIRAD UMR IATE, puis CNRS Heudiasyc), ainsi que Joël Abécassis (INRA, UMR IATE) et Bernard Cuq (Supagro, UMR IATE) sur la question applicative. La conception et l'implémentation d'outils ont été réalisés au cours des stages de Noémie Aubry (ingénieur en nutrition) co-encadré avec Bernard Cuq, et d'Iyan Johnson (ingénieur en cognitique), co-encadré avec Brigitte Charnomordic. Ils ont été poursuivis par Luc Menut (INRA UMR IATE).

La méthode de raisonnement par cas proposée a la spécificité de s'appuyer sur la notion de réconciliation de références. Habituellement les méthodes de réconciliation de références ont pour objectif de détecter que deux données différentes se réfèrent à la même entité du monde réel. Dans cette étude il ne s'agit pas de détecter des données redondantes, mais des données similaires, dans le sens où elles sont issues d'un même scénario expérimental ; des variations peuvent exister, néanmoins ces données peuvent être considérées comme résultant d'un seul et même "cas" expérimental. Cette étude a été faite avec Fatiha Saïs et implémentée par Luc Menut.

Troisième thème : Argumentation pour l'aide à la décision

Par ailleurs, de nouvelles questions ont émergé de l'analyse des enjeux au sein d'une filière. En effet, la maîtrise de la qualité au sein des filières repose sur de nombreux critères (qualité environnementale, économique, fonctionnelle, sanitaire, etc.). Les objectifs de qualité s'appuient sur différents acteurs, techniciens, gestionnaires, associations de professionnels, utilisateurs, collectivités publiques, etc. Les buts des différents acteurs d'une filière pouvant être divergents, la résolution de problèmes d'arbitrage se pose en vue de la prise de décision. Celle-ci peut se construire sur le mode du compromis (solution satisfaisant, au moins partiellement, tous les acteurs), ou privilégier certains acteurs, en fonction des priorités du décideur. Cette problématique d'arbitrage, novatrice pour l'analyse de filière, pose également des problèmes fondamentaux. Les méthodes d'arbitrage s'appuient sur les travaux en argumentation et en décision. Elles visent à mettre en place un système de décision argumentée permettant l'analyse des enjeux d'une filière et la recherche de solutions. En particulier, les mécanismes d'argumentation, qui permettent d'introduire des éléments d'explication dans la prise de décision, sont peu abordés dans la littérature concernant la décision multicritère. La conduite de ce travail a été amorcée par l'étude de la *représentation de points de vue dans une ontologie*, puis dans des connaissances représentées par des *graphes conceptuels*. Il a ensuite été développé dans le cadre de la thèse de Jean-Rémi Bourguet (soutenue fin 2010), apportant les résultats suivants :

- la définition d'un cadre formel pour la décision multicritère argumentée multi-agents ;
- l'application à un cas d'étude concernant la recommandation du PNNS de favoriser un pain de consommation courante plus complet ;
- une représentation dans le modèle des graphes conceptuels.

Ces travaux ont été engagés en partenariat avec Leïla Amgoud et Henri Prade de l'équipe ADRIA de l'IRIT, qui fait référence dans ce domaine, et en collaboration avec Marie-Laure Mugnier (LIRMM, GraphIK). Ils ont donné lieu à plusieurs publications impliquant également Jérôme Fortin et Madalina Croitoru (LIRMM, GraphIK), Joël Abécassis et Patrice Buche (INRA, UMR IATE). Une partie des résultats a été implémentée au cours du stage d'Ahmed Chadli (master 2 en informatique) que j'ai encadré. J'ai également assuré la responsabilité scientifique, sous la direction de Marie-Laure Mugnier et en collaboration avec Leila Amgoud (IRIT), de la thèse de Jean-Rémi Bourguet "Contribution aux méthodes d'argumentation pour la prise de décision.

Application à l'arbitrage au sein de la filière céréalière". Cette thèse a été soutenue en décembre 2010.

1.3 Organisation du mémoire

Ce mémoire présente un choix de travaux de recherche. La problématique de l'aide à la décision pour les filières agroalimentaires est déclinée autour de quatre contributions.

Le *chapitre 2* traite de la coopération de connaissances hétérogènes : des dires d'experts et des données expérimentales. Il présente la génération d'une ontologie ainsi qu'un mécanisme de confrontation entre les deux types de connaissances. Le *chapitre 3* présente une approche collaborative et itérative pour concevoir des modèles prédictifs pertinents. Elle associe une ontologie, une méthode d'apprentissage (arbres de décision en l'occurrence) et des retours de la part des experts. Le *chapitre 4* décrit un modèle de décision argumentée appliqué à l'analyse d'une polémique dans une politique de santé publique. Le *chapitre 5* propose une méthode d'aide à la décision en ingénierie inverse, c'est-à-dire guidée par les objectifs en aval de la filière. Elle est illustrée dans le cas de la filière boulangère.

Ces travaux permettent de dégager des perspectives qui sont discutées en conclusion dans le *chapitre 6*. Le détail des publications, des projets, des encadrements et autres responsabilités collectives est donné dans les annexes.

Chapitre 2

Intégration de formalismes hétérogènes

Ce travail se situe dans le contexte général de la construction et de la validation de l'expertise d'un domaine. Il vise la coopération de deux types de connaissances, hétérogènes par leur niveau de granularité et par leur formalisme : des dires d'experts représentés dans le modèle des graphes conceptuels et des données expérimentales représentées dans le modèle relationnel. Nous proposons d'automatiser deux étapes : d'une part, la génération d'une ontologie simple (partie terminologique du modèle des graphes conceptuels) guidée à la fois par le schéma relationnel et par les données qu'il contient ; d'autre part, l'évaluation de la validité des dires d'experts au sein des données expérimentales. La méthode que nous introduisons pour cela est fondée sur l'utilisation de graphes conceptuels patrons annotés.

Ces résultats ont été implémentés au sein d'une application concrète concernant le contrôle de la qualité alimentaire. Ils ont été publiés dans (Thomopoulos et collab., 2007, 2008). La méthode a été définie en collaboration avec Jean-François Baget et Ollivier Haemmerlé. Le cas applicatif a été étudié avec Joël Abécassis et Bernard Cuq. L'implémentation a été réalisée par Clément Molla, Amine Lakhoua et Samira Rezgane.

2.1 Problématique

La coopération de connaissances hétérogènes a été très étudiée sous un aspect particulier : l'intégration de sources hétérogènes, coopérant pour répondre à une requête de l'utilisateur, chaque source étant en mesure de fournir une partie des réponses ou encore des réponses partielles. Elle continue à être une problématique essentielle, notamment dans le cadre de la mise en correspondance d'ontologies, du fait du nombre croissant de sources d'informations disponibles via le Web. La problématique qui nous intéresse ici est toutefois différente. En effet, alors qu'en intégration de sources hétérogènes les différentes sources d'information ont le même rôle (la mise à disposition d'information en vue de répondre à une requête), ici *les différents types de connaissances n'ont pas le même statut* : une des sources contient des connaissances synthétiques, d'un niveau de granularité général et considérées comme appréhendables par l'humain, elle fournit des règles génériques sans couvrir tous les cas particuliers possibles ; les autres sources, au contraire, sont d'un niveau de granularité très fin, précises et fiables, mais trop circonstanciées pour être directement exploitables par l'humain.

Dans cette étude, les formalismes utilisés pour les différentes sources sont eux aussi hétérogènes, adaptés au type de connaissance représenté :

1. des dires d'experts, connaissances à caractère générique, découlant de l'expérience des spécialistes du domaine et décrivant les mécanismes communément admis régissant ce domaine. Ces connaissances sont représentées sous la forme de règles dans le modèle des graphes conceptuels. Nous développons dans ce chapitre la justification du choix de ce modèle de représentation des connaissances ;
2. des données expérimentales, issues de la littérature internationale du domaine. Elles sont représentées dans le modèle relationnel. Ces données nombreuses décrivent avec précision et de façon chiffrée des expériences réalisées pour approfondir la connaissance du domaine et leurs résultats. Ces résultats peuvent ou non vérifier les connaissances apportées par les dires d'experts.

La coopération des deux types de connaissances permet de tester la validité des dires d'experts sur les données expérimentales, et à plus long terme de consolider l'expertise du domaine.

Deux différences importantes entre les deux formalismes, ayant des ré-

percussions sur les vocabulaires utilisés, sont, d'une part, que les graphes conceptuels représentent des connaissances d'un caractère beaucoup plus générique que celles de la base de données relationnelle, d'autre part, que le modèle des graphes conceptuels comporte une partie ontologique (vocabulaire hiérarchisé constituant le support du modèle) contrairement au modèle relationnel. Nous proposons dans un premier temps la génération d'une ontologie, guidée par les informations de structure et les données du modèle relationnel, qui en l'occurrence préexistent aux connaissances exprimées sous forme de graphes conceptuels. Les difficultés rencontrées sont les suivantes : comment identifier, au sein du schéma relationnel et/ou des données qu'il contient, les concepts que l'on peut considérer comme pertinents pour un niveau de granularité plus général, celui des dires d'experts ? Comment hiérarchiser les différents concepts identifiés, alors que le modèle relationnel ne prend pas explicitement en compte la relation "sorte de" ? Peut-on aller plus loin dans la suggestion de concepts complémentaires pertinents ? La méthodologie proposée est semi-automatique, elle nécessite une validation experte.

Dans un deuxième temps, nous introduisons un processus permettant de tester la validité des dires d'experts au sein des données expérimentales, c'est-à-dire de réaliser l'*interrogation d'une base de données relationnelle par un système dans le formalisme des graphes conceptuels*. Cette étape est automatique. Outre la définition de l'évaluation de la validité des dires d'experts, le problème posé est celui de l'automatisation de la construction de requêtes SQL à partir de graphes conceptuels dont la forme et le contenu peuvent varier. Le processus que nous proposons s'appuie sur l'utilisation de graphes conceptuels patrons annotés.

Ce travail est illustré par une application concrète dans le domaine de la qualité alimentaire mené par l'INRA (Institut National de la Recherche Agronomique) de Montpellier.

Le chapitre est organisé de la façon suivante. La partie 2.2 rappelle un certain nombre de notions préliminaires concernant le modèle des graphes conceptuels. La partie 2.3 décrit la génération d'une ontologie, guidée par les informations de structure et les données du modèle relationnel. La partie 2.4 présente la méthode d'évaluation de la validité des dires d'experts au sein des données expérimentales. La partie 2.5 est consacrée à l'application des résultats au sein d'un projet concernant le contrôle de la qualité alimentaire. Enfin la partie 2.6 conclut et présente quelques perspectives.

2.2 Notions préliminaires

Nous rappelons ici la syntaxe et la sémantique de deux formalismes de la famille des graphes conceptuels (Sowa, 1984) : les graphes conceptuels simples et leur extension aux règles. La formalisation adoptée ici est proche de celle de (Mugnier, 2000), que le lecteur pourra consulter pour plus de précisions.

Le choix de ce formalisme pour modéliser les connaissances d'experts est justifié par les considérations suivantes, développées dans (Bos et collab., 1997) et (Genest, 2000) :

- l'aspect graphique (diagrammatique) des connaissances représentées rend la modélisation plus simple par l'expert, et son apprentissage du langage plus rapide ;
- les raisonnements sont calculés par des opérations de graphes et sont donc, eux aussi, représentables graphiquement, ce qui permet à l'expert d'affiner sa modélisation en visualisant de façon intuitive les conséquences de celle-ci.

2.2.1 Les graphes conceptuels simples

Les graphes conceptuels simples forment un langage correspondant au fragment positif, conjonctif, existentiel de la logique du premier ordre. Il a été introduit (Sowa, 1976) comme une interface graphique pour les bases de données relationnelles.

Syntaxe Dans ce langage, un vocabulaire encode les connaissances du niveau ontologique (des noms de classes et leur hiérarchie), tandis que les graphes encodent des connaissances factuelles (les instances et les relations entre elles).

Définition 2.1 *Un* vocabulaire *est un n-uplet* $\mathcal{V} = ((T_C, \leq_C), (T_1, \leq_1), \ldots, (T_k, \leq_k))$ *d'ensembles finis, partiellement ordonnés et deux à deux disjoints où les éléments de* T_C *sont des* types de concepts *et les éléments de* T_i *sont des* types de relations *d'arité* i*. Nous nous donnons également deux ensembles disjoints* M *et* V *de* marqueurs individuels *et de* noms de variables.

Définition 2.2 *Un graphe conceptuel simple défini sur un vocabulaire* $\mathcal{V}$ *est un quintuplet* $G = (C, R, \gamma, \tau, \mu)$ *où* C *est un ensemble de* concepts, R *est*

un ensemble de relations, $\gamma : R \to C^+$ *associe à chaque relation un tuple de concepts (ses* arguments*), dont la taille est le* degré *de la relation ;* τ *associe à chaque concept de* C *un élément de* T_C *et à chaque relation de degré* i *de* R *un élément de* T_i *(leur type) ;* μ *associe à chaque concept* c *de* C *un marqueur individuel de* M *(*c *est dit individuel) ou un nom de variable de* V *(*c *est dit générique).*

Un vocabulaire est représenté par les diagrammes de Hasse de ses ordres partiels. Nous représentons un graphe simple de la façon suivante : chaque concept c est représenté par un rectangle à l'intérieur duquel est inscrite la chaîne $\tau(c) : \mu(c)$; chaque relation r est représentée par un ovale contenant la chaîne $\tau(r)$; si c est le i-ième argument de la relation r, on dessine un trait entre les représentations de c et r, et on inscrit i à côté de ce trait. Ainsi, la figure 2.1 représente le graphe simple défini par : $G = (C, R, \gamma, \epsilon_C, \epsilon_R)$ où : $C = \{c_1, c_2, c_3, c_4\}; R = \{r_1, r_2, r_3\}$; $\gamma(r_1) = (c_1, c_2)$, $\gamma(r_2) = (c_1, c_3)$, $\gamma(r_3) = (c_3, c_4)$; $\tau(c_1) =$ Aliment, $\tau(c_2) =$ Cuisson à l'eau, $\tau(c_3) =$ Vitamine, $\tau(c_4) =$ Teneur, $\tau(r_1) =$ subit, $\tau(r_2) =$ contient, $\tau(r_3) =$ caractérisé ; $\mu(c_1) =$ Frekeh, $\mu(c_2) = x1$, $\mu(c_3) = x2$, $\mu(c_4) = x3$ (où Frekeh est un marqueur individuel, et $x1, x2, x3$ sont des noms de variables).

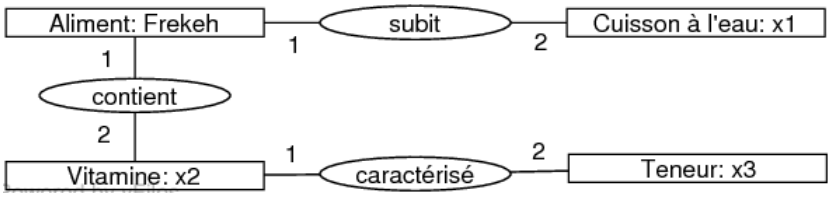

FIGURE 2.1 – Un graphe conceptuel simple G

Sémantique L'opérateur Φ associe une formule logique à un vocabulaire ou à un graphe simple. Le problème de déduction entre graphes simples peut ainsi être défini par le problème de déduction des formules logiques associées. Ces formules sont obtenues de la façon suivante :

Interprétation d'un vocabulaire Soient t et t' deux types de relations d'arité i. On note $\phi(t, t') = \forall x_1 \ldots \forall x_i (t(x_1, \ldots, x_i) \to t'(x_1, \ldots, x_i))$. L'interprétation du vocabulaire $\Phi(\mathcal{V})$ est la conjonction, pour toute paire de types

telle que $t \leq t'$, des formules $\phi(t, t')$. Notons que les types de concepts sont interprétés comme des types de relations d'arité 1.

Interprétation d'un graphe A chaque concept c nous associons l'atome $\phi(c) = \tau(c)(\mu(c))$, où $\mu(c)$ est une constante si c est individuel, une variable sinon ; et à chaque relation r telle que $\gamma(r) = (c_1, \ldots, c_i)$, l'atome $\tau(r)(\mu(c_1), \ldots, \mu(c_i))$. Notons $\phi(G)$ la conjonction des $\phi(x)$ pour tous les concepts et relations de G. Alors $\Phi(G)$ est la fermeture existentielle de $\phi(G)$.

Par exemple, la traduction par Φ du graphe de la figure 2.1, représentant l'information "le frekeh subit une cuisson à l'eau et contient une vitamine caractérisée par une certaine teneur", est : $\exists x1\ \exists x2\ \exists x3$ (Aliment(Frekeh) $\wedge$ Cuisson à l'eau($x1$) $\wedge$ Vitamine($x2$) $\wedge$ Teneur($x3$) $\wedge$ subit(Frekeh, $x1$) $\wedge$ contient(Frekeh, $x2$) $\wedge$ caractérisé($x2$, $x3$)).

Le problème d'inférence dans les graphes conceptuels simples consiste à savoir si on peut déduire un graphe Q (répondre à la requête Q) à partir d'une base de connaissances constituée d'un graphe G, ou d'un ensemble de graphes (la base de faits) et d'un vocabulaire.

Définition 2.3 *Soient G et Q deux graphes simples définis sur un vocabulaire $\mathcal{V}$. On dit que Q est conséquence de G et on note $G \models Q$ ssi $\Phi(\mathcal{V})$, $\Phi(G) \models \Phi(Q)$.*

Inférences Le calcul de conséquence entre graphes simples est efficacement réalisé par une sorte d'homomorphisme de graphes étiquetés appelé projection.

Définition 2.4 *Soient $G = (C_G, R_G, \gamma_G, \tau_G, \mu_G))$ et $Q = (C_Q, R_Q, \gamma_Q, \tau_Q, \mu_Q)$ deux graphes simples définis sur un vocabulaire$\mathcal{V}$. Une* projection *de Q dans G est une application π de C_Q dans C_G telle que :*

- $\forall c, c' \in C_Q, \mu_Q(c) = \mu_Q(c') \Rightarrow \pi(c) = \pi(c')$;
- $\forall c \in C_Q$, *c est individuel* $\Rightarrow \mu_G(\pi(c)) = \mu_Q(c)$;
- $\forall c \in C_Q, \tau_G(\pi(c)) \leq_C \tau_Q(c)$;
- $\forall r \in R_Q$, *avec* $\gamma(r) = (c_1, \ldots, c_k)$, $\exists r' \in R_G$ *tq* $\gamma(r') = (\pi(c_1), \ldots, \pi(c_k))$ *et* $\tau(r') \leq_k \tau(r)$.

Théorème 2.1 *(Mugnier, 2000) Soient G et Q deux graphes simples définis sur un vocabulaire $\mathcal{V}$, où G est sous forme normale*[1]*. Il existe une projection*

1. Un graphe simple est sous forme normale quand tous ses concepts ont un marqueur distinct. Tout graphe peut être transformé en un graphe normal équivalent.

de Q dans G ssi $G \models Q$.

2.2.2 Règles de graphes conceptuels

Syntaxe Les règles (Salvat, 1998) forment une extension des graphes conceptuels dans laquelle on ajoute à une base de connaissance des règles de la forme "si A alors B" où A et B sont deux graphes simples. L'ajout de règles augmente fortement l'expressivité du langage.

Définition 2.5 *Une règle (de graphe conceptuel simple) définie sur un vocabulaire $\mathcal{V}$ est une paire $R = (H, C)$ de graphes simples définis sur $\mathcal{V}$. H est l'hypothèse de R et C sa conclusion.*

Une règle est représentée graphiquement par les deux graphes qui la composent, séparés par un symbole d'implication allant de l'hypothèse vers la conclusion, comme dans la figure 2.2 qui représente la règle : "si un aliment subit une cuisson à l'eau et contient une vitamine caractérisée par une certaine teneur, alors cette teneur montre une diminution".

FIGURE 2.2 – Une règle de graphe conceptuel simple R

Sémantique L'opérateur Φ est étendu afin de traduire les règles. Si $R = (H, C)$ est une règle, alors $\Phi(R) = \forall x_1 \ldots \forall x_i(\phi(H) \rightarrow (\exists y_1 \ldots \exists y_j \phi(C)))$, où les x_p sont les noms de variables de H et les y_q sont les noms de variables de C qui ne sont pas dans H.

La formule logique associée à la règle R de la figure 2.2 est $\Phi(R) = \forall x1\ \forall x2\ \forall x3\ \forall x4\ ((\text{Aliment(x1)} \wedge \text{Cuisson à l'eau}(x2) \wedge \text{Vitamine}(x3) \wedge \text{Teneur}(x4) \wedge \text{subit(x1, } x2) \wedge \text{contient(x1, } x3) \wedge \text{caractérisé}(x3, x4)) \rightarrow (\exists y1\ (\text{Teneur}(x4) \wedge \text{Diminution}(y1) \wedge \text{montre}(x4, y1))))$.

Définition 2.6 *Soient G et Q deux graphes simples définis sur un vocabulaire $\mathcal{V}$ et $\mathcal{R} = \{R_1, \ldots, R_k\}$ un ensemble de règles définies sur $\mathcal{V}$. On dit que*

Q est conséquence de G et $\mathcal{R}$ et on note $G, \mathcal{R} \models Q$ ssi $\Phi(\mathcal{V})$, $\Phi(G)$, $\Phi(R_1)$, ..., $\Phi(R_k) \models \Phi(Q)$.

Inférences Le calcul de déduction en présence de règles peut se faire en marche avant ou en marche arrière (voir Baget et Salvat (2006) pour une présentation de ces deux méthodes). Nous présentons ici brièvement la marche avant.

Définition 2.7 *Une règle $R = (H, C)$ est dite applicable à un graphe simple G s'il existe une projection π de H dans G. Dans ce cas, appliquer R à G suivant π consiste à faire l'union disjointe*[2] *de G et de $sp(\pi, C)$, où $sp(\pi, C)$ est obtenu en remplaçant le marqueur de tout concept c de C identique à celui d'un sommet c' de H par le marqueur de $\pi(c')$; puis à mettre sous forme normale le graphe obtenu.*

Théorème 2.2 *(Salvat, 1998) Soient G et Q deux graphes simples définis sur un vocabulaire $\mathcal{V}$, et $\mathcal{R}$ un ensemble de règles définies sur $\mathcal{V}$. Il existe une séquence finie d'applications de règles de $\mathcal{R}$ qui transforme G en un graphe simple G' tq $G' \models Q$ ssi $G, \mathcal{R} \models Q$.*

2.3 Génération d'une ontologie

Nous nous situons dans le cas où un recueil de données expérimentales détaillées représentées dans le modèle relationnel préexiste à l'expression de connaissances expertes d'un niveau de granularité plus général. L'objectif est d'automatiser autant que possible la génération d'une ontologie simple, constituant l'ensemble des types de concepts de la partie terminologique du modèle des graphes conceptuels, à l'aide du schéma et des données relationnels existants.

Dans cette partie, après une présentation de travaux proches, nous décrivons trois étapes de la génération de l'ontologie : l'identification de types de concepts de haut niveau, la hiérarchisation de ces types de concepts, la proposition de types de concepts complémentaires.

2. L'union disjointe de deux graphes est le graphe dont le dessin est la juxtaposition de leur dessin.

2.3.1 Travaux proches

Cette problématique nécessitant une expertise importante, une méthode totalement automatisée pour la génération d'une ontologie (Pernelle et collab., 2001) est exclue. Notre objectif est différent de l'apprentissage de concepts telle qu'abordée par les approches FCA (Formal Concept Analysis, voir Tilley et collab. (2005) ou ILP (Inductive Logic Programming, voir Muggleton et Raedt (1994)), qui s'appuient sur l'existence de propriétés communes à des sous-ensembles de données pour les regrouper en de nouveaux concepts. Ici l'objectif premier est d'identifier et de hiérarchiser des concepts pertinents pour l'expression de connaissances expertes, parmi ceux déjà présents dans les données de façon peu explicite et avec une structure inappropriée.

La recherche d'une structure hiérarchique, en particulier d'une structure arborescente, dans des données semi-structurées (Termier et collab., 2002) ou non structurées (Kietz et collab., 2000; Folch et collab., 2004) a été étudiée, notamment dans le cadre relativement récent de l'échange et de l'interrogation de données sur le Web. En revanche, la recherche d'une nouvelle structure pour des objectifs spécifiques dans des données déjà structurées, qui est le but visé ici, est peu courante. Des travaux proches sont ceux qui touchent la question de la cohabitation entre vocabulaires hétérogènes, tels que la transformation de modèles (Sendall et Kozaczynski, 2003) et l'alignement d'ontologies (Euzenat et collab., 2004). En alignement d'ontologies, des correspondances sont établies entre des vocabulaires préexistants, conçus indépendamment les uns des autres, tandis que dans cette étude l'ontologie est dérivée des données.

Des graphes conceptuels aux bases de données La correspondance entre graphes conceptuels simples et requêtes conjonctives en bases de données est bien connue (Kolaitis et Vardi, 1998; Mugnier, 2000). Soit $\mathcal{V}$ un vocabulaire, et G et Q deux graphes simples sur $\mathcal{V}$. G et Q sont transformés (en G' et Q') de la façon suivante : les types de concepts sont transformés en relations unaires, et chaque concept de type t devient un concept sans type, incident à une relation unaire typée t. Pour chaque relation r de type t, pour chaque supertype t' de t, nous rajoutons une nouvelle relation r' de type t' telle que $\gamma(r) = \gamma(r')$. Il s'ensuit que $G \models_{\mathcal{V}} Q$ ssi $\Phi(G') \models \Phi(Q')$ (nous n'avons plus besoin des formules traduisant le support, toutes leurs conséquences ont été traduites dans les graphes). Puisque $\Phi(G')$ et $\Phi(Q')$

sont des formules positives conjonctives, nous pouvons définir $\mathcal{B}$ comme les tables ayant comme formule logique associée $\Phi(G')$ et A comme la requête ayant comme formule logique associée $\Phi(Q')$. Nous avons ainsi $G \models_{\mathcal{V}} Q$ ssi il existe une réponse à A dans $\mathcal{B}$. Cependant, cette correspondance repose sur une identification entre le vocabulaire des graphes conceptuels et le schéma de bases de données, hypothèse trop forte comme nous le verrons par la suite.

Le système Sym'Previus (Haemmerlé et Carbonneill, 1996) proposent d'ajouter une couche "graphes conceptuels" à une base de données relationnelle (BDR) préexistante. Cette couche sert d'interface en vue de permettre une complétion des requêtes se fondant sur la sémantique des attributs présents dans la BDR. Chaque attribut de la BDR est intégré à l'ensemble des types de concepts (sous des types de concepts génériques qui spécifient le type de la donnée au sens SQL du terme). D'autres types de concepts sont ajoutés manuellement à cet ensemble afin de disposer de connaissances supplémentaires qui sont exploitées par spécialisation ou généralisation au moment de l'expression des requêtes dans le modèle des graphes conceptuels.

Le système Sym'Previus a été développé dans le cadre d'un projet de recherche français sur un outil de prévention du risque microbiologique dans les aliments (Haemmerlé et collab., 2006). Cet outil repose sur trois bases distinctes, ajoutées au système successivement au fur et à mesure du développement du projet : une base de données relationnelle, une base de graphes conceptuels et une base de données XML. Les trois bases sont interrogées simultanément et uniformément par le biais d'une interface unique, qui se fonde sur une même ontologie.

Cette ontologie a été construite manuellement, au moment de l'ajout de la base de graphes conceptuels au système. Un schéma de base de données relationnelle ainsi que ses données préexistaient. L'ensemble des attributs correspondant à des entités significatives de l'application a été partitionné en deux : les attributs dont les valeurs pouvaient être hiérarchisées selon la relation "sorte de" (substrat, germe pathogène...) et les attributs dont les valeurs étaient des ensembles intrinsèquement "plats" (les noms d'auteurs de publications, par exemple). Tous les noms d'attributs significatifs ont été ajoutés à l'ontologie Sym'Previus en tant que types de concepts. Les valeurs apparaissant dans les colonnes correspondant à des attributs à valeurs hiérarchisées ont été insérés en tant que sous-types de concepts dans l'ontologie. Leur positionnement précis dans la hiérarchie a été réalisé manuellement par

les experts.

2.3.2 Identification de types de concepts de haut niveau

Dans cette étape, l'objectif est d'identifier des types de concepts de haut niveau (niveau de granularité général). Nous identifions deux types d'entités, que nous considérons comme susceptibles de correspondre à des types de concepts de haut niveau pertinents :

- celles dont les occurrences portent un nom, c'est-à-dire qui ont un attribut "nom" (ou encore "intitulé", "libellé", contenant la chaîne "nom", etc.). Nous supposons en effet que ces entités sont de caractère plus général, par opposition aux entités secondaires correspondant à des informations plus détaillées, dont les occurrences ne sont pas nommées mais identifiées uniquement par des identifiants numériques. Ce sont les premières qui sont utiles pour l'expression des dires d'experts : ceux-ci manipulent des notions désignées par un nom et non des informations circonstantielles détaillées ;
- celles qui peuvent être subdivisées en sous-catégories. Nous cherchons pour cela les entités qui ont un attribut "catégorie" (ou encore "famille", "type", etc.). Nous supposons en effet que ces entités, du fait de la classification engendrée par leurs sous-catégories, fournissent des types de concepts pertinents pour l'ontologie.

La frontière entre les deux cas n'est pas absolue et est très dépendante du type de modélisation. Par exemple, l'attribut "nom" d'une entité peut parfaitement avoir pour valeurs des sous-catégories de l'entité considérée. Ainsi dans le cas de notre application, l'entité *Constituant nutritionnel* a un attribut "nom" destiné à prendre des valeurs telles que "Vitamine", "Lipide", etc., qui ne désignent pas à proprement parler des instances, mais des familles de constituants nutritionnels. Si l'on prend un exemple très courant sortant du cadre de notre application, une entité *Personne* ayant un attribut "nom" peut cacher des utilisations différentes : le plus souvent, le nom d'une personne ("Dupont" par exemple) désigne un individu particulier (même si elle n'en est pas l'identifiant) ; mais si l'on se situe dans le contexte d'une application en généalogie, "Dupont" peut désigner une branche d'individus.

Du fait de cette proximité, les deux cas seront traités de façon homogène par la suite. Par souci de simplification, nous n'indiquerons pas systématiquement la liste complète des attributs considérés ("nom", "catégorie", "famille", etc.) mais nous les désignerons sous le terme d'*attributs indicateurs*. Ces at-

tributs sont de type chaîne de caractères.

Définition 2.8 *On appelle* attribut indicateur *tout attribut dont le nom figure dans une liste prédéfinie de termes déclarés propres à exprimer la dénomination ou la classification. Un tel attribut est considéré comme appartenant à une entité d'un niveau de granularité général.*

Remarque 2.1 *Ce processus permet de proposer des types de concepts de haut niveau pertinents. Etant donnée la variabilité des modélisations, il nécessite une vérification experte.*

Utilisation du schéma relationnel Dans un premier temps, nous nous appuyons sur le schéma de la base de données relationnelle. D'un point de vue ingénierie des bases de données, après une modélisation à l'aide par exemple du modèle entité-association, on sait qu'une relation (ou table) du schéma de la base de données relationnelle correspond :

- soit à une entité du domaine représenté. Elle en comporte alors les attributs. Elle peut également comporter les identifiants d'autres entités (avec lesquelles elle était liée par une association), plus rarement des attributs d'association ;
- soit à une association (de type plusieurs à plusieurs) entre entités. Elle comporte alors comme attributs leurs identifiants et les attributs d'association.

La table obtenue porte généralement le nom de l'entité ou de l'association correspondante.

Afin d'identifier les types de concepts de haut niveau, nous faisons les hypothèses simplificatrices suivantes :

1. les entités, plutôt que les relations, véhiculent les principaux concepts du domaine représenté. Les types de concepts de haut niveau sont donc à rechercher dans les noms d'entités, autrement dit parmi les noms de tables du schéma relationnel ;
2. le cas d'une association ayant un attribut indicateur est considéré comme exceptionnel.

Définition 2.9 *Sont considérés comme* types de concepts de haut niveau issus du schéma relationnel *les noms des tables qui comportent (au moins)*

un attribut indicateur. Les types des concepts de haut niveau ainsi identifiés sont ajoutés à l'ontologie.

Exemple 2.1 *Dans le cas de notre application, des exemples de types de concepts de haut niveau issus du schéma sont les suivants :* Aliment, Changement, Constituant, Méthode, Opération, Propriété, Variable, ... *En revanche, d'autres comme* Valeur par défaut, Valeur expérimentale, *etc., n'ont pas été considérés comme des types de concepts de haut niveau.*

Utilisation des données relationnelles Dans un second temps, nous nous intéressons aux valeurs prises par les attributs indicateurs. Nous avons fait l'hypothèse que les attributs indicateurs sont susceptibles de prendre pour valeurs des sous-catégories de l'entité à laquelle ils appartiennent. La prise en compte des données relationnelles permet par conséquent de proposer comme types de concepts de haut niveau les valeurs des attributs indicateurs. Leur organisation hiérarchique est précisée dans la partie 2.3.3.

Définition 2.10 *Sont considérés comme* types de concepts de haut niveau issus des données *les valeurs prises par les attributs indicateurs de la base de données. Les types des concepts de haut niveau ainsi identifiés sont ajoutés à l'ontologie.*

Exemple 2.2 *Dans le cas de notre application, ont par exemple été définis comme types de concepts de haut niveau issus des données les types de concepts suivants :* Augmentation, Diminution, Protéine, Lipide, Vitamine, Vitamine B, Qualité, Teneur, ...

2.3.3 Hiérarchisation des types de concepts

Deux niveaux de hiérarchisation sont proposés :

• la hiérarchisation des types de concepts de haut niveau issus des données par rapport à ceux issus du schéma : la valeur prise par un attribut indicateur d'une table (type de concept de haut niveau issu des données) est considérée comme une spécialisation du type de concept portant le nom de cette table (type de concept de haut niveau issu du schéma). Par exemple, *Vitamine* est une spécialisation de *Constituant nutritionnel*;

• la hiérarchisation des types de concepts de haut niveau issus des données entre eux : elle s'appuie sur l'inclusion des labels des types de concepts. Par exemple, *Vitamine B* (désignant la famille des vitamines B) est une spécialisation de *Vitamine*.

La définition 2.11 résume les étapes 2.3.2 et 2.3.3, leur résultat est soumis à vérification experte.

Définition 2.11 *La* génération d'une ontologie simple *O à partir de la base de données relationnelle est réalisée de la façon suivante. Pour chaque table, de nom noté* T*, de la base de données, si la table* T *comporte au moins un attribut indicateur, alors :*

• *le type de concept (de haut niveau issu du schéma relationnel)* T *est ajouté à* O *;*

• *pour chaque attribut indicateur de* T*, prenant un ensemble de valeurs* $v_1, \ldots, v_n$ *:*

– *le type de concept (de haut niveau issu des données)* v_i*, sous-type de* T*, est ajouté ;*

– *si* v_i *est inclus dans* v_j $(i, j \in [1, n])$*, alors* v_j *est un sous-type de* v_i*.*

Exemple 2.3 *Par exemple dans le cas de notre application la table* Constituant *comporte l'attribut indicateur* nom_constituant*, prenant pour valeurs* Protéine, Lipide, Vitamine, *etc.*

Le type de concept (de haut niveau issu du schéma relationnel) Constituant *et les types de concepts (de haut niveau issus des données)* Protéine, Lipide, Vitamine, Vitamine B *sont ajoutés à* O *comme sous-types de* Constituant. *"Vitamine" étant inclus dans "Vitamine B", le type de concept* Vitamine B *est sous-type de* Vitamine *(voir figure 2.3).*

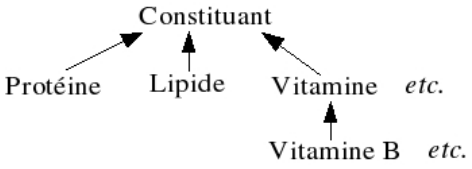

FIGURE 2.3 – Exemple de hiérarchisation des types de concepts

2.3.4 Proposition de types de concepts complémentaires

La méthode proposée dans cette partie afin de compléter l'ontologie par la suggestion de types de concepts supplémentaires pertinents, est spécifique à la forme des connaissances expertes considérée dans l'application. Nous nous situons dans le cas suivant. Les connaissances expertes sont exprimées par des règles de la forme "si (hypothèse) alors (conclusion)". Plus précisément, il s'agit de règles de causalité exprimant une relation de cause à effet entre (i) un ensemble de conditions, décrit par l'hypothèse, interagissant entre elles pour produire (ii) l'effet qui en résulte, décrit par la conclusion.

Par exemple, une règle experte simple issue de l'application est la suivante : "si un aliment, caractérisé par une teneur en vitamines, subit une cuisson à l'eau, alors cette teneur diminue". Elle est représentée par la règle de graphes conceptuels de la figure 2.2.

La nature des interactions existant entre les concepts apparaissant dans l'hypothèse n'est pas toujours bien connue des experts. En particulier, ces interactions peuvent être dues à l'interférence d'autres concepts qui ne sont pas nécessairement identifiés et explicités. L'objectif de cette partie est de mettre en évidence certains de ces concepts. La méthode proposée est fondée sur la comparaison de descriptions textuelles des concepts apparaissant dans l'hypothèse.

En effet, les tables de la base de données relationnelle qui ont permis d'obtenir les types de concepts apparaissant dans l'hypothèse (cf. définition 2.11) fournissent parfois des descriptions textuelles, contenues dans la valeur d'un attribut nommé par exemple "description", "commentaires", etc. Pour chaque paire de types de concepts apparaissant dans une même hypothèse de règle experte et pour lesquelles une telle description est disponible, la démarche proposée consiste à rechercher dans ces descriptions l'existence de termes communs.

Exemple 2.4 *La comparaison des decriptions textuelles de certaines opérations (Cuisson à l'eau, Cuisson vapeur, Hydratation, Séchage) avec les descriptions textuelles de certains constituants (Son de blé, Fibre, Lipide, Vitamine, Polyphénol) ont en commun le terme "eau". En effet, ces opérations unitaires ont toutes un effet sur la teneur en eau (apport ou retrait d'eau) et ces constituants possèdent tous des sous-catégories ayant une affinité particulière avec l'eau (solubilité ou absorption particulières). La mise en évidence du terme commun "eau" a conduit les experts à compléter, d'une part, par*

l'ajout du type de concept Eau*, d'autre part, par la spécialisation de types de concepts existants pour faire apparaître des catégories ayant une interaction particulière avec l'eau : ainsi* Vitamine *est spécialisé en* Vitamine hydrosoluble *(surtype, entre autres, de* Vitamine B*, qui est soluble dans l'eau) et* Vitamine liposoluble.

Les résultats obtenus sont nombreux et doivent être triés manuellement par l'expert.

La recherche de termes communs fait appel à des techniques de traitement de la langue naturelle, en particulier la suppression des mots creux ("stop-words"), l'homogénéisation des variations syntaxiques (tokenisation, lemmatisation).

2.4 Evaluation de la validité des dires d'experts

Contrairement à la partie précédente (section 2.3) qui nécessite une intervention experte, la méthode présentée dans cette partie est automatique. L'objectif est de tester si les connaissances expertes exprimées sous forme de règles de graphes conceptuels sont valides au sein des données expérimentales de la base relationnelle. Un taux de validité de la règle testée est calculé – similaire au support dans les règles d'association – et les données faisant exception à la règle sont identifiées et visualisées par l'utilisateur.

Dans cette partie, après une présentation des travaux existants, nous définissons ce que nous entendons par l'évaluation de la validité d'une règle, introduisons les notions de patron et d'instance de règle, enfin exposons le déroulement de la validation d'une instance de règle.

2.4.1 Problématiques proches

On peut distinguer deux formes de cohabitation entre une base de données relationnelle et une base de connaissances dans le modèle des graphes conceptuels :

- il n'y a pas d'échange de données entre les deux modèles, en revanche ceux-ci sont exploités en utilisant un formalisme commun (pivot) pour l'expression des requêtes et/ou de l'ontologie du domaine. Le projet Sym'Previus (Haemmerlé et collab., 2006) est un exemple d'application où le forma-

lisme pivot est un langage de requêtes inspiré du formalisme relationnel. Le cas inverse (interrogation d'une BD par des requêtes graphes conceptuels) est celui qui nous intéresse ici ;

• il y a échange de données entre les deux modèles. Ce cas se rencontre par exemple : (i) s'il y a nécessité de migration de données vers l'un des deux formalismes jouant le rôle d'entrepôt. Ce cas a été envisagé, mais pas exploré, comme perspective au projet Sym'Previus, où le modèle des graphes conceptuels est utilisé comme formalisme de stockage provisoire et souple de données non prévues par le schéma relationnel ; (ii) si l'un des deux formalismes paraît plus adapté pour la résolution de certains types de problèmes, et que l'on fait le choix d'utiliser le formalisme le plus adapté à les traiter. Ce cas n'a pas fait l'objet de travaux à notre connaissance.

2.4.2 Calcul du taux de validité

Evaluer la validité d'une règle experte au sein des données expérimentales consiste à calculer la proportion de données satisfaisant à la fois l'hypothèse et la conclusion de cette règle, parmi celles qui en satisfont l'hypothèse. Si l'on note n_H le nombre de données satisfaisant l'hypothèse et $n_{H\wedge C}$ le nombre de données satisfaisant à la fois l'hypothèse et la conclusion, le taux de validité V d'une règle est $V = \frac{n_{H\wedge C}}{n_H} \times 100$, où n_H et $n_{H\wedge C}$ sont le résultat de requêtes SQL effectuant un comptage (*select count*) des données remplissant respectivement les critères de satisfaction de l'hypothèse et les critères de satisfaction de l'hypothèse et de la conclusion. Le problème qui se pose est celui de l'automatisation de la construction de ces requêtes.

2.4.3 Notions de patron de règle, d'instance de règle et propriétés associées

Bien que les règles expertes puissent prendre des formes variables, il est possible de les regrouper en ensembles de règles qui suivent la même forme générale.

Exemple 2.5 *Les règles expertes représentées par les figures 2.2 et 2.4 sont de la même forme.*

La "forme générale" d'un ensemble de règles expertes peut elle-même être représentée par une règle, appelée patron de règle. Sa structure est identique

FIGURE 2.4 – Exemple de règle experte de même forme que celle de la figure 2.2

à celle des règles expertes de cet ensemble, mais ses sommets concepts sont plus généraux que ceux des règles expertes de l'ensemble. Autrement dit, chacune des règles expertes de l'ensemble a un graphe hypothèse et un graphe conclusion qui sont des spécialisations (par restriction des étiquettes) de ceux du patron de règle. Ces règles sont appelées instances de règle. Les graphes hypothèse et conclusion du patron de règle se projettent donc dans ceux de chacune de ses instances.

Exemple 2.6 *Les règles des figures 2.2 et 2.4 sont des instances du patron de règle de la figure 2.5.*

FIGURE 2.5 – Exemple de patron de règle

Le niveau de généralité des types de concepts utilisés dans un patron de règle n'est pas quelconque : il s'agit de concepts de haut niveau issus du schéma relationnel. Au contraire, les types de concepts utilisés dans une instance de règle peuvent être des concepts de haut niveau issus des données (les marqueurs peuvent de plus être individuels). Cette particularité est essentielle pour le déroulement de la validation d'une instance de règle.

Définition 2.12 *Un* patron de règle *est une règle, dans le formalisme des graphes conceptuels, dont les concepts ont pour types des types de concepts de haut niveau issus du schéma relationnel et dont les marqueurs sont génériques. Une* instance de règle *est une règle, dans le formalisme des graphes*

conceptuels, obtenue par restriction des étiquettes des sommets concepts d'un patron de règle donné. L'instance de règle est dite conforme *à ce patron.*

En conséquence, les types de concepts apparaissant dans un patron de règle fournissent une liste de noms de tables de la base de données (les concepts de haut niveau issus du schéma). L'hypothèse (respectivement, la conclusion) d'un patron de règle peut être interprétée, au sein de la base de données, comme la formule d'une requête permettant de sélectionner les données satisfaisant l'hypothèse (respectivement, la conclusion). Cette formule fait intervenir les tables apparaissant comme types de concepts dans l'hypothèse (respectivement, la conclusion) du patron de règle. Cette formule ne fait que spécifier un schéma de requête. Elle n'est pas contrainte pas des critères de sélection particuliers. De tels critères n'apparaîtront que lors du traitement des instances de règles, présenté en 2.4.4.

Définition 2.13 *Soit H l'hypothèse d'un patron de règle. Soit Q une requête sur la base de données relationnelle permettant de sélectionner les données satisfaisant H. Q s'écrit en termes de calcul relationnel sous la forme $\{T|F(T)\}$, où F est une formule, T une variable n-uplet de F et $F(T)$ une évaluation de F. La réponse à la requête Q sera un ensemble de n-uplets $\{t|F(t)$ vraie$\}$. F est construite par la conjonction des formules suivantes.*

- Formules atomiques associées aux concepts de H : *Soit $s_{c_1}, \ldots, s_{c_n}$ les concepts de H, de types $c_1, \ldots, c_n$ (ce sont des types de concepts de haut niveau issus du schéma relationnel et donc des tables de la base de données relationnelle). Les concepts de H étant génériques, chaque concept s_{c_i} fournit la formule atomique : $\exists x_i, c_i(x_i)$.*
- Formules associées aux relations de H : *Soit s_r un sommet relation de H avec $\gamma(s_r) = (s_{c_k}, \ldots, s_{c_l})$. Deux cas de figure peuvent se présenter :*

– *le schéma de Q ne fait pas intervenir d'autres tables que celles présentes dans H pour joindre les tables $c_k, \ldots, c_l$. Chaque concept s_{c_k}, ..., s_{c_l} de $\gamma(s_r)$ fournit au moins une formule atomique*[3] *de la forme : $x_i.a_i = X_i$, où a_i désigne un attribut de la table c_i et X_i une constante ou une expression $x_j.a_j$ ($j \in [k, l]$, a_j attribut de c_j).*

– *le schéma de la requête Q fait intervenir d'autres tables que celles présentes dans H pour joindre les tables $c_k, \ldots, c_l$. Soit $t_m, \ldots, t_p$ ces tables.*

3. Ces formules atomiques ne sont pas nécessairement distinctes de celles fournies par les autres voisins de s_r, par exemple un voisin peut fournir $x_i.a_i = x_j.a_j$ et un autre $x_j.a_j = x_i.a_i$.

Chacune d'entre elles fournit une formule atomique $\exists x_i, t_i(x_i)$ *et au moins une formule atomique* $x_i.a_i = X_i$. *Le sommet relation* s_r *fournit alors une formule (non-atomique) de la forme :* $\exists x_m, \ldots, x_p, t_m(x_m) \wedge \ldots \wedge t_p(x_p) \wedge x_k.a_k = X_k \wedge \ldots \wedge x_l.a_l = X_l \wedge x_m.a_m = X_m \wedge \ldots x_p.a_p = X_p$.

- Attributs recherchés : *Soit* $attr_1, \ldots, attr_q$ *les attributs recherchés, issus respectivement des tables* $tbl_1, \ldots, tbl_q$ *(*$attr_i$ *non nécessairement distinct de* a_j, $j \in [k,l] \cup [m,p]$ *et* tbl_i *dans* $\{c_1, \ldots, c_n, t_m, \ldots, t_p\}$*). F(t) est contrainte par :* $t.attr_i = tbl_i.attr_i$ *(*$i \in [1,q]$*).*

Dans le cas général, $F(t)$ *est donc de la forme :* $\exists x_1, \ldots, x_n, x_m, \ldots, x_p, c_1(x_1) \wedge \ldots \wedge c_n(x_n) \wedge t_m(x_m) \wedge \ldots \wedge t_p(x_p) \wedge x_k.a_k = X_k \wedge \ldots \wedge x_l.a_l = X_l \wedge x_m.a_m = X_m \wedge \ldots x_p.a_p = X_p \wedge t.attr_1 = tbl_1.attr_1 \ldots t.attr_z = tbl_q.attr_q$.

A ce stade (patrons de règle), cette formule ne peut être qu'en partie générée de façon automatique. En effet, les tables $t_m, \ldots, t_p$, les attributs a_i et les termes X_i ne peuvent pas toujours être calculés. Les limites de l'automatisation sont dues à l'ambiguïté des jointures entre tables et du fait des possibilités multiples que l'on peut rencontrer en cas de jointures intermédiaires à réaliser. La formule F doit donc être définie par le concepteur pour l'hypothèse de chaque patron de règle.

La formule de la requête permettant de sélectionner les données satisfaisant la conclusion d'un patron de règle est construite de la même façon. Enfin, la formule de la requête permettant de sélectionner les données satisfaisant à la fois l'hypothèse et la conclusion d'un patron de règle est obtenue par conjonction des formules associées à l'hypothèse et à la conclusion.

Pour permettre l'évaluation d'une règle experte (voir partie 2.4.2), les deux requêtes nécessaires sont celle comptant les données satisfaisant l'hypothèse et celle comptant les données satisfaisant à la fois l'hypothèse et la conclusion d'un patron de règle. Ces deux requêtes sont associées manuellement par le concepteur à chaque patron de règle.

Exemple 2.7 *La formule associée à l'hypothèse du patron de règle de la figure 2.5 est*

$\exists x_1, x_2, x_3, x_4, x_5, x_6, aliment(x_1) \wedge operation(x_2) \wedge constituant(x_3) \wedge propriete(x_4) \wedge resultat(x_5) \wedge etude(x_6) \wedge x_1.id\_aliment = x_5.id\_aliment \wedge x_2.id\_operation = x_6.id\_operation \wedge x_3.id\_constituant = x_5.id\_sous\_constituant \wedge x_4.id\_propriete$

$= x_5.id\_\ propriete \wedge x6.id\_etude = x5.id\_etude \wedge t.x_4.id\_resultat = x_5.id\_resultat$

La requête SQL associée à l'hypothèse du patron de la figure 2.5 est

```
SELECT COUNT(resultat.id_resultat)
FROM resultat, aliment, constituant, etude, operation
WHERE resultat.id_aliment = aliment.id_aliment
AND etude.id_operation = operation.id_ operation
AND resultat.id_sous_constituant = constituant.id_constituant
AND resultat.id_ propriete = propriete.id_propriete
AND resultat.id_etude = etude.id_etude
```

A chaque concept d'un patron de règle est associé une information, destinée à indiquer la spécialisation de ce sommet concept au sein des instances de règle conformes à ce patron :

- si le type de concept de ce sommet a des sous-types (concepts de haut niveau issus des données), de quel attribut de la table sont-ils des valeurs ? Cet attribut est supposé le même pour toutes les instances d'un patron de règle donné ;
- si le marqueur de ce sommet est susceptible d'être individuel au sein des instances de règle, de quel attribut de la table ces marqueurs sont-ils des valeurs ? On suppose l'existence d'un tel attribut et, là encore, cet attribut est supposé le même pour toutes les instances.

Exemple 2.8 *Dans le patron de règle de la figure 2.5, le type* Constituant *peut être spécialisé par des sous-types qui sont aussi des valeurs de l'attribut* nom_constituant *de la table* Constituant.

Ainsi dans les figures 2.2 et 2.4, Vitamine *et* Minéral*, qui sont des spécialisations du type de concept* Constituant*, sont aussi des valeurs de l'attribut* Constituant.nom_constituant.

Définition 2.14 *Un patron* annoté *est un patron de règle P auquel sont associés :*

- *une requête d'hypothèse, dénombrant le nombre de n-uplets de la base de données satisfaisant l'hypothèse de P ;*
- *une requête d'hypothèse et conclusion, dénombrant le nombre de n-uplets de la base de données satisfaisant à la fois l'hypothèse et la conclusion de P ;*

• *pour chacun de ses sommets concepts s_c (de type c), deux attributs :*

– *un attribut de type, indiquant l'attribut de la table c contenant les spécialisations (notées c'_i) du type de concept c attendues (le cas échéant), dans les instances de règle conformes à P, pour un sommet image de s_c (par l'opération de projection) ;*

– *un attribut de marqueur, indiquant l'attribut de la table c contenant les marqueurs des types de concept c ou c'_i attendus (le cas échéant), dans les instances de règle conformes à P, pour un sommet image de s_c (par l'opération de projection).*

Remarque 2.2 *Les formules des requêtes associées à un patron de règle ne faisant que spécifier un schéma de requête, le résultat des deux requêtes est par définition identique, c'est-à-dire égal au nombre de données de la base. Un patron de règle a donc une validité de 100 %.*

2.4.4 Déroulement de la validation d'une instance de règle

Afin de tester la validité d'une règle experte, c'est-à-dire d'une instance de règle, ce qui est l'objectif recherché, deux nouvelles requêtes vont être construites automatiquement : une requête dénombrant les données satisfaisant l'hypothèse de l'instance de règle (appelée requête d'hypothèse) et une requête dénombrant les données satisfaisant à la fois l'hypothèse et la conclusion de l'instance de règle (appelée requête d'hypothèse et conclusion).

Ces requêtes sont composées de deux parties :

• leurs premières parties respectives décrivent le schéma de la requête à exécuter : il s'agit des requêtes associées au patron de règle auquel se conforme l'instance de règle à évaluer. Ces parties sont donc fournies par les annotations du patron de règle ;

• leurs secondes parties permettent de sélectionner les seuls n-uplets qui prennent les valeurs d'attributs correspondant aux spécialisations réalisées dans l'instance de règle. Ces parties spécifient donc des critères de sélection, qui vont être construits automatiquement en utilisant comme attributs de sélection les annotations du patron de règle (attributs de type et attributs de marqueur) et comme valeurs de sélection les types de concepts et les marqueurs présents dans l'instance de règle à évaluer.

Définition 2.15 *Soit P un patron de règle et I une instance de règle à valider, conforme à P.*

La requête d'hypothèse (respectivement d'hypothèse et conclusion) de I, notée Q_H (resp. $Q_{H \wedge C}$), est la conjonction de :

- *la requête d'hypothèse (resp. d'hypothèse et conclusion) associée à P ;*
- *l'ensemble des critères de sélection de la forme attribut = valeur obtenus comme suit. Soit π une projection de P dans I. Soit $sc = [c, m]$ un sommet concept de l'hypothèse de P (resp. de P entier) et $sc' = [c', m']$ son image dans I par π.*
  - *Si $c' < c$ (au sens de la relation de spécialisation) alors un critère de sélection est créé, ayant pour attribut l'attribut de type associé à sc et pour valeur c' (type de concept de haut niveau issu des données, qui correspond à une valeur prise par l'attribut de type associé à sc). Si de plus c' a des sous-types, dans l'ensemble des types de concepts, alors pour chacun de ces sous-types c'' un critère de sélection est créé, ayant pour attribut l'attribut de type associé à sc et pour valeur c''.*
  - *Si $m' < m$ (au sens de la relation de spécialisation) alors un critère de sélection est créé, ayant pour attribut l'attribut de marqueur associé à sc et pour valeur m'.*

Remarque 2.3 *S'il existe plusieurs projections de P dans I, une requête d'hypothèse (resp. d'hypothèse et conclusion) de I est obtenue pour chaque projection. Seule la requête d'hypothèse (resp. d'hypothèse et conclusion) donnant le plus grand résultat (nombre de données) est retenue : on estime qu'elle correspond à la spécialisation escomptée du patron de règle.*

Exemple 2.9 *La requête SQL associée à l'hypothèse de l'instance de règle de la figure 2.2.*

```
SELECT COUNT(resultat.id_resultat)
FROM resultat, aliment, constituant, etude, operation
WHERE resultat.id_aliment = aliment.id_aliment
AND etude.id_operation = operation.id_operation
AND resultat.id_sous_constituant = constituant.id_constituant
AND resultat.id_propriete = propriete.id_propriete
AND resultat.id_etude = etude.id_etude
// Partie de la requete ajoutée à celle du patron (voir Exemple 2.7)
AND operation.nom_operation = 'Cuisson à l'eau' propriete.nom_propriete = 'Teneur'
```

```
AND (constituant.nom_constituant = 'Vitamine'
// Partie correspondant à la prise en compte des sous-types de Vitamine (les
types Cuisson à l'eau et Teneur n'ont pas de sous-type dans cet exemple)
OR constituant.nom_constituant = 'Vitamine liposoluble'
OR constituant.nom_constituant = 'Vitamine E' ...)
```

Remarque 2.4 *Dans les exemples 2.7 et 2.9, la table* Etude *intervient comme table de jointure (voir définition 2.13).*

Les requêtes Q_H et $Q_{H \wedge C}$ ont respectivement pour résultat n_H et $n_{H \wedge C}$, qui permettent de calculer le taux de validité de l'instance de règle. Les règles dont le taux de validité est strictement inférieur à 100 % ont des exceptions au sein de la base de données . Ces exceptions peuvent être sélectionnées et affichées à l'utilisateur.

Exemple 2.10 *Pour la règle de la figure 2.2, on a un taux de validité V de 97.5 % (voir les figures 2.7 et 2.8).*

2.5 Application

Les méthodes présentées ont été mises en œuvre au sein d'une application concernant le contrôle de la qualité alimentaire. L'enjeu est d'améliorer la maîtrise des facteurs de contrôle de la qualité nutritionnelle. Après une présentation de l'environnement de travail, nous décrirons les données expérimentales et les connaissances expertes de l'application, puis leur validation.

2.5.1 Environnement de travail

Les données expérimentales sont regroupées au sein d'une base de données MySQL. La consultation et la saisie des données par des spécialistes du domaine se fait via un navigateur, au travers de formulaires PHP. La base de données contient à l'heure actuelle une trentaine de tables et les résultats détaillés d'environ 600 expériences.

Les règles expertes sont représentées à l'aide de l'interface graphique CoGUI (http ://www. lirmm.fr/ gutierre/cogui/). Environ 150 règles expertes sont disponibles, une vingtaine est utilisée pour tester la méthodologie proposée, à commencer par les cas les plus simples.

La communication entre les deux systèmes est établie à l'aide d'une connexion JDBC.

2.5.2 Description des données expérimentales

Destiné à des scientifiques et à des industriels de l'agroalimentaire, l'outil (en langue anglaise) de consultation et de saisie des données expérimentales intègre des données scientifiques, aussi exhaustives que possible, issues de la littérature traitant des qualités nutritionnelles des aliments à base de blé dur, et décrivant l'impact des procédés de transformation sur ces qualités. Un tel article scientifique comporte généralement les informations suivantes :

- des mesures expérimentales, notamment sur l'analyse des constituants nutritionnels (par exemple : dosage de la vitamine B1 dans les pâtes) ;
- des résultats sur l'impact d'une ou plusieurs opérations unitaires sur une ou plusieurs qualités nutritionnelles (effet de la cuisson-extrusion sur la teneur en minéraux) ;
- des données sur l'influence de certains paramètres de l'opération unitaire (par exemple : l'effet de la température de stockage sur la rétention des vitamines) ;
- des données sur l'influence d'autres opérations unitaires (par exemple : l'effet du type de séchage des pâtes sur la rétention des vitamines pendant la cuisson dans l'eau) ;
- des modèles décrivant l'évolution des qualités nutritionnelles (par exemple : la cinétique de dégradation thermique de la vitamine B1) ;
- des références bibliographiques.

2.5.3 Description des connaissances expertes

Les types de concepts du vocabulaire utilisé pour exprimer les connaissances expertes ont été obtenus comme présenté dans la partie 2.3. Dans la modélisation du support, l'essentiel de la sémantique est porté par les types de concepts. Les types de relations constituent quant à eux des connecteurs généraux aussi stables que possible. La figure 2.6 montre une partie de ce vocabulaire, créé à l'aide de CoGUI.

Les connaissances expertes, représentées par des règles de graphes conceptuels, expriment, pour chaque opération unitaire intervenant dans le process de fabrication d'un aliment à base de blé dur, et pour chaque constituant

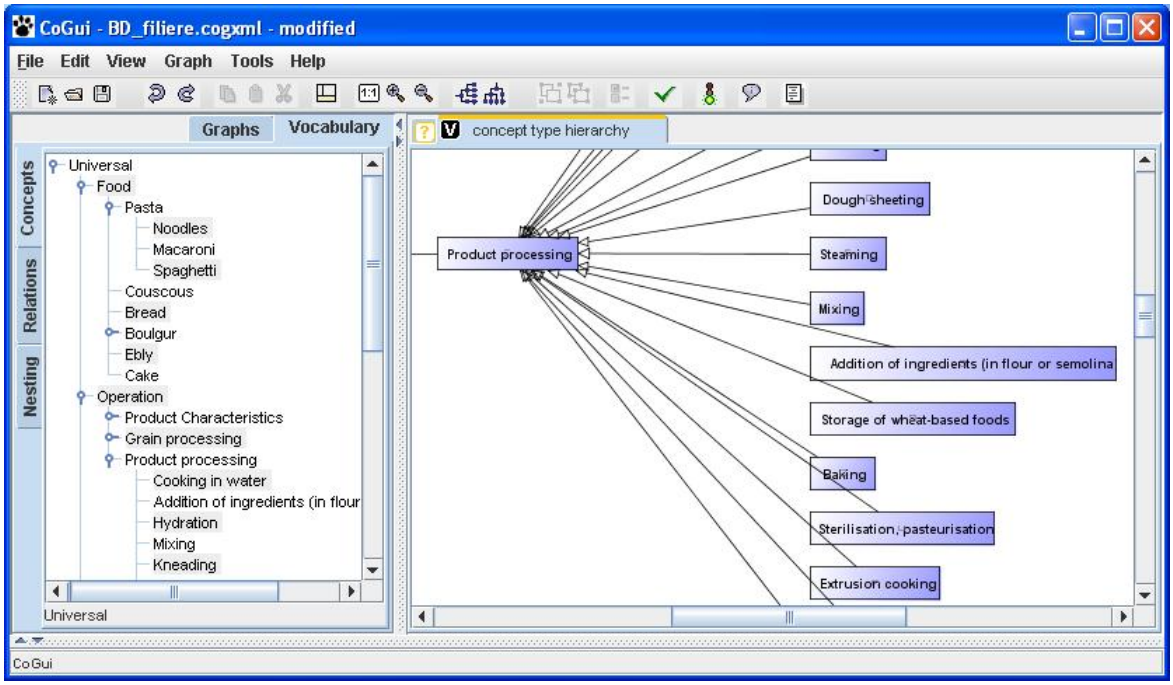

FIGURE 2.6 – Une partie du vocabulaire utilisé pour exprimer les connaissances expertes

nutritionnel répertorié, l'impact ou les impacts connu(s) de cette opération sur ce constituant, en explicitant les conditions dans lesquelles cet impact semble se produire, pouvant faire intervenir des interactions avec d'autres opérations unitaires.

L'impact peut concerner la variation de la teneur du constituant (augmentation, diminution, stagnation) mais aussi la modification de propriétés qualitatives du constituant, telles que la digestibilité, l'allergénicité, etc.

2.5.4 Validation des connaissances expertes

L'évaluation des connaissances expertes peut être visualisée de deux façons par l'utilisateur : elle peut être effectuée individuellement règle par règle, et permet alors à l'utilisateur d'obtenir l'affichage des données expérimentales faisant exception à cette règle ; elle peut également être effectuée sous la forme d'un tableau récapitulatif de l'ensemble des règles déclarées dans l'applica-

tion et de leurs taux de validité respectifs. Les figures 2.7 et 2.8 illustrent le premier cas de figure.

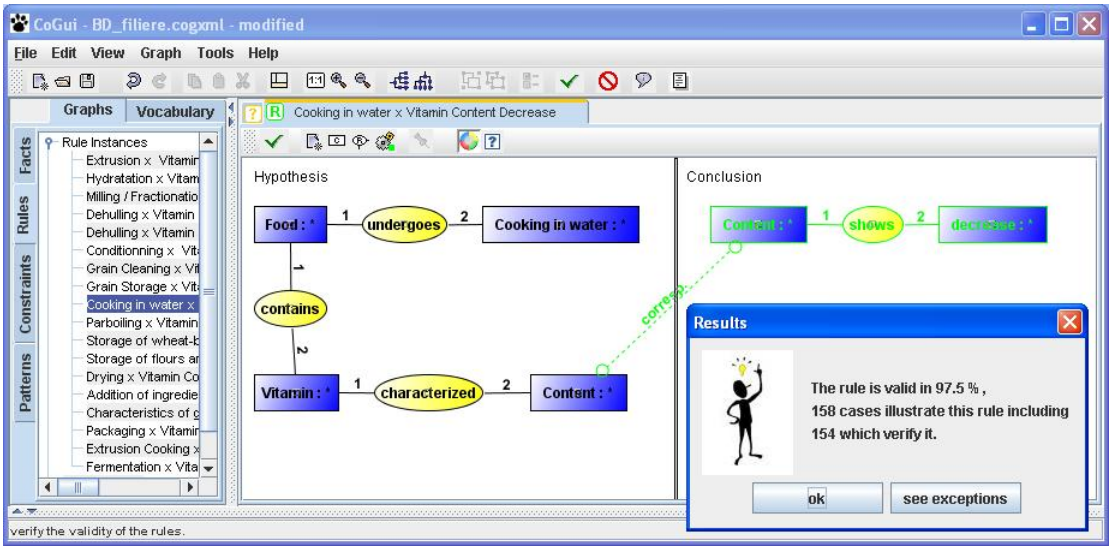

FIGURE 2.7 – Evaluation de la validité d'une règle experte

2.6 Conclusion du chapitre

Etant donnés deux types d'information hétérogènes disponibles pour un domaine (des connaissances expertes génériques exprimées par des règles de causalité d'une part, des résultats expérimentaux détaillés d'autre part) représentés dans deux formalismes distincts (respectivement le modèle des graphes conceptuels et le modèle relationnel), nous avons présenté dans ce chapitre deux étapes pour la construction d'une expertise du domaine : (i) la génération d'une ontologie par l'identification de types de concepts de haut niveau au sein du schéma relationnel et au sein des données relationnelles et la hiérarchisation de ces types de concepts. Cette étape est automatique mais soumise à vérification experte ; (ii) l'évaluation de la validité des connaissances expertes au sein des données expérimentales. Cette étape est fondée sur la notion de patron de règle dans le formalisme graphes conceptuels, auquel est associé un "squelette" de requête SQL correspondant dans le formalisme relationnel. L'évaluation d'une instance de règle donnée conforme à

Cooking in water x Vitamin Content Decrease
RULE EXCEPTIONS

Food	Parameters of the unit operation				Interactions with prior unit operations	Results				Reference
	Temperature(°C)	% of salt in water(%)	Time(mn)	Kind of water()		Name of the component	Percentage		Kind of result	
							Value	Standard deviation		
Pasta	100	Undefined	20	Undefined	Yes	Vitamin A	4%		increase	Parrish, D. B. et al.,1980
Pasta	100	Undefined	20	Undefined	Yes	Vitamin A	8%		increase	Parrish, D. B. et al.,1980
Pasta	100	Undefined	20	Undefined	Yes	Vitamin A	8%		increase	Parrish, D. B. et al.,1980

FIGURE 2.8 – Affichage des exceptions d'une règle experte

un patron, se fait alors en complétant le "squelette" associé à ce patron par des critères de sélection spécifiques à l'instance de règle considérée. Cette étape est automatique, ce qui est permis par des annotations effectuées sur les patrons de règle.

La méthodologie proposée est ainsi fondée sur la coopération des deux types d'information et des deux formalismes hétérogènes. Elle est illustrée par un cas d'application concret.

L'objectif à plus long terme des règles de causalité est une utilisation à des fins d'aide à la décision : étant donnée une requête de l'utilisateur, exprimant un but souhaité, la question est de déterminer quelles conditions permettent d'atteindre ce but, en recherchant des règles dont la conclusion satisferait le but souhaité, et dont l'hypothèse fournirait des conditions suffisantes à son obtention. Cette question sera reprise dans le chapitre 5.

Chapitre 3

Méthodes prédictives

Les expériences scientifiques sont productrices de données qui peuvent alimenter les connaissances sur un domaine. Cependant, les données issues d'expériences ne sont pas nécessairement formalisées dans un objectif de découverte de connaissances. Une approche collaborative entre les connaissances expertes et les connaissances issues de l'exploration de données est un bon moyen de tirer parti de ces données. Une telle approche nécessite une formalisation des connaissances du domaine dans le but de guider l'acquisition des connaissances à travers les données. Ce chapitre présente une approche itérative, guidée par les données et s'appuyant sur une ontologie, pour concevoir des modèles pertinents. Elle associe une ontologie modélisant les connaissances du domaine et une méthode d'apprentissage pour construire des modèles interprétables (arbres de décision dans ce chapitre). Une évaluation subjective et objective est impliquée dans le processus. Une étude de cas dans le domaine de l'agroalimentaire montre l'intérêt de cette méthode. L'approche a été menée en collaboration étroite entre quatre chercheurs en informatique (Brigitte Charnomordic, Sébastien Destercke, Iyan Johnson et moi-même) et deux chercheurs en sciences de l'aliment (Joël Abécassis et Bernard Cuq), avec une implication régulière de tous les participants. Elle est présentée plus en détail dans (Johnson et collab., 2010; Thomopoulos et collab., 2013a).

3.1 Problématique

Le partage de l'expertise et l'apprentissage à partir des données sont deux piliers de la construction d'outils d'aide à la décision efficaces, en particulier dans des domaines sans modèle mathématique établi. La première clé de succès de tels modèles est leur fiabilité. Pour être utilisés par les experts, ils doivent produire des résultats sûrs. Même si la confiance dans ces modèles peut être partiellement obtenue par une procédure de validation, les modèles *interprétables* s'avèrent mieux obtenir l'adhésion et la confiance des experts, car il permettent de comprendre le cheminement du raisonnement au-delà de la prédiction. Des exemples de modèles d'apprentissage interprétables sont les arbres de décision, les bases de règles floues (Guillaume et Charnomordic, 2011), les réseaux bayésiens, etc.

Un deuxième élément-clé (lié au premier) concerne la pertinence des jeux de données utilisés par les méthodes d'apprentissage. Partir du présupposé que les données sont idéalement structurées et pertinentes pour l'objectif visé est une hypothèse très forte, souvent insatisfaite, en particulier lorsque les données proviennent de diverses sources et de dispositifs expérimentaux différents. Dans le domaine des sciences du vivant, comme dans d'autres domaines portant sur des systèmes complexes, les données expérimentales sont souvent dédiées à l'étude d'une question scientifique précise et ne sont pas conçues avec une visée plus globale. Cela est dû au fait que couvrir tous les aspects du système nécessiterait des expérimentations très coûteuses. Rien, dans les données brutes, ne garantit donc que les variables seront pertinentes lorsqu'utilisées dans un modèle d'apprentissage conçu pour un objectif plus large. Dans (Vialette et collab., 2005), une approche statistique est appliquée pour traiter cette question, fondée sur la méta-analyse de résultats de prédiction obtenus à partir d'un assemblage de données provenant de différentes sources. Cependant, cette approche nécessite d'avoir préalablement validé un modèle prédictif, par exemple dans (Vialette et collab., 2005), une équation modélisant la croissance bactérienne. Dans le présent chapitre, nous considérons le cas plus déroutant où aucun modèle validé n'est disponible. Les entretiens avec des experts sont alors fondamentaux pour évaluer la pertinence des modèles construits. Toutefois, ce sont aussi des tâches coûteuses en temps. Les chercheurs en intelligence artificielle qui se sont intéressés à l'élicitation de connaissances qualitatives détenues par les experts connaissent les difficultés de cette tâche (Gaines et Shaw, 1993). En conséquence, les inter-

ventions expertes doivent rester limitées, et être guidées pour être efficaces.

Notre contribution consiste en une méthode de construction de modèles guidés par les données :

- collaborative, puisqu'elle fait interagir méthodes d'apprentissage et experts ;
- itérative, car elle implique plusieurs cycles pour améliorer les résultats obtenus ;
- hybride, du fait qu'elle s'appuie à la fois sur les données et les connaissances.

L'approche proposée vise trois objectifs connexes : (i) trouver des variables explicatives pertinentes, (ii) structurer et enrichir les connaissances du domaine sous la forme d'une ontologie, (iii) accroître la confiance des experts dans le modèle. Son principe est le suivant. Partant d'un ensemble de données initial et de connaissances représentées dans une ontologie, un premier modèle guidé par les données est appris (étape 1). Ce modèle est d'abord évalué avec des critères numériques objectifs. Puis il est soumis aux experts du domaine, qui peuvent enrichir l'ontologie en suggérant de nouvelles relations entre certaines variables (étape 2). Des transformations sont appliquées aux données selon ces nouvelles relations (étape 3). L'objectif est de remplacer les variables jugées non pertinentes par les experts par des variables significatives. L'ensemble du processus est répété de façon itérative jusqu'à ce qu'aucune possibilité d'amélioration ne soit plus détectable, aboutissant potentiellement à un modèle de prédiction pointu, conforme aux connaissances expertes et aux données. La figure 3.1 décrit ce processus.

Le chapitre est organisé comme suit. La partie 3.2 propose un état de l'art. La partie 3.3 est consacrée à la définition de l'ontologie et à son interaction avec les données. La partie 3.4 décrit les différentes opérations de traitement des données effectuées à l'aide de l'ontologie. La partie 3.5 explique le principe de l'approche proposée. Une étude de cas concernant l'impact des procédés de transformation agroalimentaires sur la qualité nutritionnelle de produits à base de blé est présentée dans la partie 3.6.

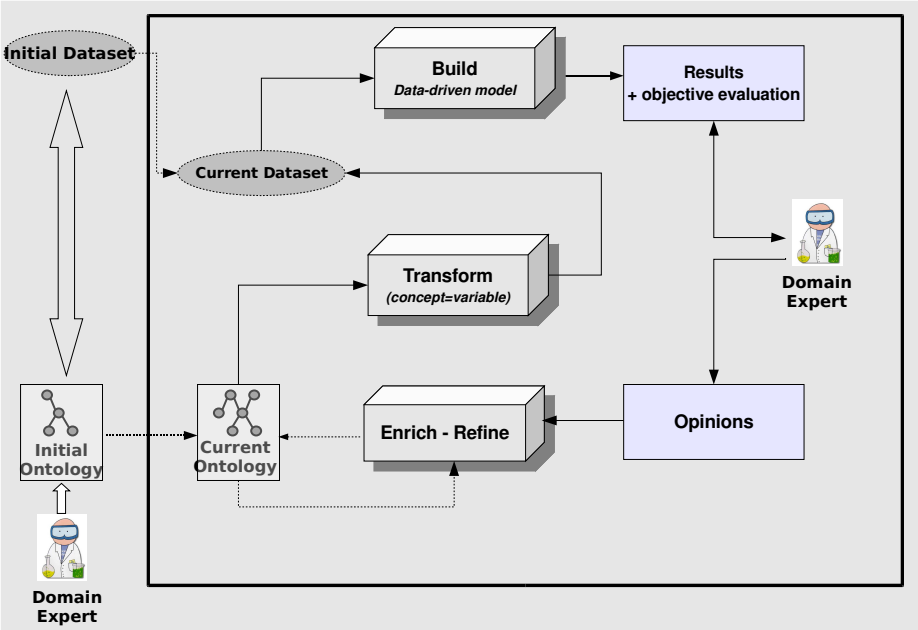

FIGURE 3.1 – Schéma du processus de construction de modèle

3.2 Littérature pertinente

3.2.1 Utilisation d'ontologies pour guider l'apprentissage

En apprentissage à partir de données, distinguer les motifs significatifs de ceux qui sont inutiles est une tâche délicate. Il semble donc naturel d'utiliser les connaissances disponibles pour faciliter la reconnaissance de tels motifs. Cependant, les propositions de méthodes génériques combinant ontologies et modèles guidés par les données sont encore peu nombreuses, notamment dans les sciences expérimentales et en sciences du vivant en particulier. Dans ce dernier domaine, la plupart des approches se situent dans le champ de la bioinformatique (Popescu et Xu, 2009) avec des préoccupations très spécifiques concernant la biomédecine et la génomique.

Quelques tentatives concernent des cas où les données sont bien structurées, ce qui facilite l'automatisation. Un exemple est la classification d'images, avec l'utilisation de concepts visuels (Maillot et Thonnat, 2008). D'autres travaux sont axés sur des problèmes où le passage à l'échelle est une question essentielle, et où les performances de la méthode peuvent être mesurées automatiquement par une évaluation numérique. Dans les domaines où de grandes quantités de données doivent être traitées, tels que le web sémantique (Stumme et collab., 2006), des approches collaboratives semi-automatiques permettant de guider le processus d'exploration ont récemment été appliquées. Une proposition d'évaluation et d'enrichissement d'ontologie est faite dans (Parekh et Gwo, 2004), en utilisant plusieurs ontologies combinées avec une approche de fouille de textes et de glossaires spécifiques du domaine. Des algorithmes ainsi qu'un outil logiciel pour la découverte collaborative à partir de sources d'informations sémantiquement hétérogènes sont décrits dans (Caragea et collab., 2005).

Le cas de l'apprentissage inductif utilisant des ontologies, des données et des arbres de décision a été abordé dans (Zhang et collab., 2002). Toutefois, il est limité au cas spécifique des taxinomies [1]. Des études similaires sont appliquées à des classifieurs bayésiens dans (Zhang et collab., 2006).

1. Ontologies arborescentes.

3.2.2 Utilisation de l'analyse subjective pour la sélection de règles ou de données

Il existe également des cas hors sciences expérimentales où l'utilisation complètement automatisée des connaissances ontologiques s'avère difficile, et où la participation d'experts semble inévitable pour améliorer les résultats des méthode guidées par les données. Dans de tels cas, l'utilisation de modèles interprétables par les experts est essentielle.

Par exemple, (Adomavicius et Tuzhilin, 2001) propose une validation guidée par l'expert de groupes de règles, pour faciliter leur validation dans des modèles à base de règles. (Ling et collab., 2008) propose une approche pour la sélection ou la fouille de données efficaces, applicable à des domaines où les entrepôts de données sont trop diffus et les connaissances disponibles trop complexes pour être directement utilisables. L'approche implique largements les experts humains, profitant de leur expertise à chaque étape, et utilise des techniques de fouille de données en assistance, pour la découverte de tendances dans les données et pour la vérification des conclusions des experts. Dans un document récent (Mansingh et collab., 2011), les auteurs se concentrent sur l'utilisation d'ontologies pour faciliter le post-traitement, par les experts du domaine, de règles d'association. Ils proposent une méthode hybride d'élagage utilisant l'analyse objective et subjective.

Notons que la plupart de ces méthodes ne sont pas fondées sur une procédure de retour sur le résultat. Deux cas peuvent être distingués :

- les connaissances ontologiques ou les connaissances expertes sont utilisées pour améliorer les résultas de l'apprentissage guidé par les données ;
- l'apprentissage guidé par les données est utilisé pour identifier d'éventuels éléments de connaissance ontologique ou pour compléter les connaissances expertes.

Les deux tâches, à savoir acquérir les connaissances ontologiques et construire des modèles interprétables ayant la confiance des experts à partir des données, sont généralement difficiles à atteindre et souffrent de certaines limitations. Il semble donc utile de construire des méthodes visant à combiner les points forts de chacune.

3.2.3 Les arbres de décision comme modèles interprétables

Les algorithmes d'arbres de décision sont des approches efficaces pour la découverte de relations complexes et non triviales guidée par les données. Leur lisibilité et l'absence d'hypothèses *a priori* expliquent leur popularité. Ils sont particulièrement utiles pour la sélection de variables dans des problèmes multidimensionnels, et de ce fait idéaux pour présenter des variables statistiquement significatives sur lesquelles les experts pourront se concentrer. Les arbres de décision peuvent être élagués pour limiter leur complexité, comme discuté dans (Ben-David et Sterling, 2006). Cette faible complexité est essentielle pour l'interprétabilité du modèle (Miller, 1956). Ce sont donc de bons candidats pour des méthodes où les experts doivent interagir avec les modèles guidés par les données.

Toutefois, les modèles guidés par les données sont fortement tributaires des données disponibles et de la méthode d'apprentissage. Tout en étant à même de produire de bonnes prédictions, il n'est pas exclu qu'ils utilisent pour cela des variables considérées par l'expert comme impropres à décrire le phénomène considéré. Différentes causes sont possibles : ces variables peuvent s'avérer être marginalement impliquées dans le phénomène, ou devenir significatives dans un certain contexte, ou en combinaison avec d'autres variables.

Dans ce chapitre, nous nous concentrons sur les arbres de décision de type C4.5 (Quinlan, 1993), que nous utilisons en classification.

Description de l'algorithme

Les arbres de classification prennent en entrée un jeu d'exemples d'entraînement, ayant chacun un ensemble de valeurs correspondant à des variables d'entrée X_k, qui peuvent être continues ou discrètes, et une variable de sortie discrète Y divisée en M_Y classes. L'objectif est d'apprendre à partir des exemples d'entraînement une structure récursive (prenant la forme d'une arborescence) comportant (i) des noeuds feuilles étiquetés par le nom d'une classe et (ii) des noeuds intermédiaires de test (chacun associé à une variable donnée) qui peuvent avoir deux issues ou plus, chacune associée à un sous-arbre.

Les inconvénients connus des arbres de décision sont leur sensibilité aux données extrêmes et le risque de surapprentissage. Pour éviter le surappren-

tissage, une validation croisée est incluse dans la procédure et pour gagner en robustesse, une étape d'élagage suit généralement l'étape de construction de l'arbre (Quinlan, 1986, 1993).

Critères de segmentation

Nous désignons par $p_m(S)$ la proportion d'exemples au noeud S qui appartiennent à la classe m ($m \in [1; M_Y]$). Pour sélectionner la variable de segmentation au noeud S, l'algorithme C4.5 examine tour à tour toutes les variables candidates et calcule le gain potentiel apporté par chacune d'elles. Il sélectionne alors la variable qui maximise le gain.

Notons M_k le nombre de modalités de X_k. Le gain acquis en segmentant le noeud S en M_k sous-ensembles $S_1, S_2 \ldots S_{M_k}$ selon X_k est évalué comme :

$$G(S, X_k) = I(S) - \sum_{i=1}^{M_k} \frac{|S_i|}{|S|} I(S_i)$$

où M_k est le nombre d'issues possibles. $I(S)$ est dérivé de l'entropie en théorie de l'information, sa valeur au noeud S est :

$$I(S) = -\sum_{m=1}^{M_Y} p_m(S) log_2\ p_m(S).$$

3.3 Définition de l'ontologie en lien avec les données

Nous considérons la description d'un domaine composée de deux éléments :

- un ensemble de données, sous la forme d'une liste d'expériences et des valeurs qu'elles prennent sur un ensemble de variables ;
- une ontologie contenant les connaissances terminologiques du domaine.

L'ontologie Ω est définie comme un couple $\Omega = (\mathcal{C}, \mathcal{R})$ où $\mathcal{C}$ est un ensemble de concepts et $\mathcal{R}$ un ensemble de relations. Un exemple d'ontologie concernant les procédés alimentaires est présenté sur la figure 3.2 (les flèches correspondent à la relation de subsomption de la partie 3.3.3). Nous utiliserons cette ontologie comme exemple illustratif par la suite.

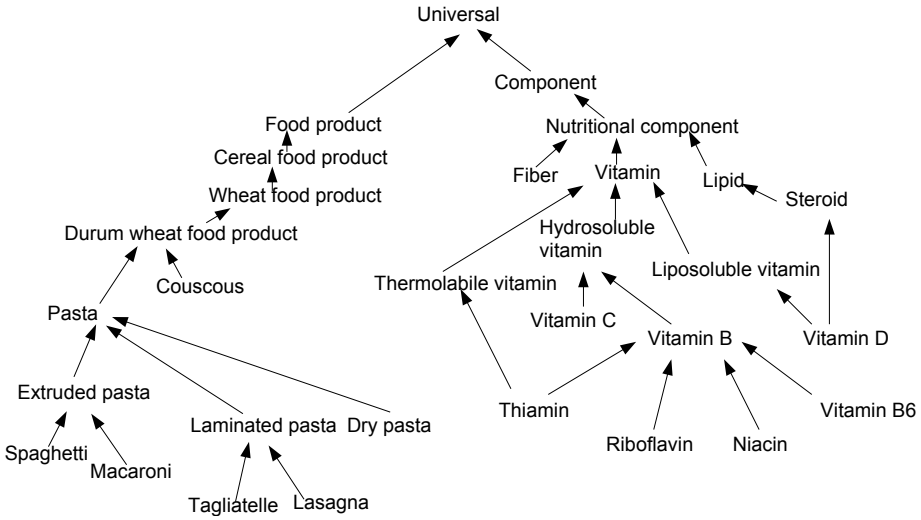

FIGURE 3.2 – Un extrait de l'ontologie utilisée pour les procédés alimentaires

3.3.1 Domaine de définition des concepts

Un concept c est associé à un domaine de définition par la fonction *Range*. Ce domaine de définition peut être :

- *numérique*, c'est-à-dire $Range(c)$ est un intervalle fermé $[min_c, max_c]$;
- *symbolique hiérarchisé*, c'est-à-dire $Range(c)$ est un ensemble d'éléments partiellement ordonnés, qui sont eux-mêmes des concepts appartenant à $\mathcal{C}$.

Le domaine associé à un concept c par la fonction *Range* correspond à l'espace sur lequel une variable associée prend ses valeurs : si on associe au concept c la variable X_k, alors $Range(c)$ est l'espace sur lequel X_k prend ses valeurs.

3.3.2 Relation entre concepts et variables

Nous considérons un ensemble de données provenant de l'observation effective de N expériences, avec K variables. Chaque ligne représente une instance d'expérience. Nous supposons que chaque variable X_k, $k = 1, \ldots, K$ est un concept $c \in \mathcal{C}$ de l'ontologie Ω. La n^{eme} valeur de la k^{eme} variable est notée $x_{k,n}$, comme l'illustre le tableau 3.1, $x_{k,n}$ appartient à $Range(X_k)$, symbolique ou numérique. Ainsi l'ensemble des concepts $\mathcal{C}$ inclut l'ensemble des variables utilisées pour la description des données.

observation	X_1	X_2	...	X_K
exp_1	$x_{1,1}$	$x_{2,1}$	...	$x_{K,1}$
exp_2	$x_{1,2}$	$x_{2,2}$	...	$x_{K,2}$
...	...	...	...	...
exp_N	$x_{1,N}$	$x_{2,N}$	...	$x_{K,N}$

TABLE 3.1 – Ensemble de données expérimentales où $\{X_k\}, k \in [1, K]$ est l'ensemble des variables et $\{exp_n\}, n \in [1, N]$ est l'ensemble des observations à partir des expériences.

3.3.3 L'ensemble des relations

L'ensemble des relations $\mathcal{R}$ est constitué de :

1. la relation de *subsomption*, également appelée relation "sorte de" et notée $\preceq$, qui définit un ordre partiel sur $\mathcal{C}$. Etant donné un concept $c \in \mathcal{C}$, nous notons $\mathcal{C}_c$ l'ensemble des sous-concepts de c, tel que :

$$\mathcal{C}_c = \{c' \in \mathcal{C} | c' \preceq c\}. \tag{3.1}$$

   Lorsque c représente une variable dont le domaine de définition est symbolique hiérarchisé, on a $Range(c) = \mathcal{C}_c$. Par souci de concision, nous utilisons par la suite $\mathcal{C}_c$ autant que possible. Par exemple, dans la figure 3.2 et pour $c = Laminated\ pasta$, on a $\mathcal{C}_{Laminated\ pasta} = \{Tagliatelle, Lasagna\}$;

2. un ensemble de *dépendances fonctionnelles*. Une dépendance fonctionnelle FD exprime une contrainte entre deux ensembles de variables et est représentée par une relation entre deux ensembles de concepts de $\mathcal{C}$. On dit qu'un ensemble de concepts $X = \{X_{k_1}, \ldots, X_{k_2}\} \subseteq \mathcal{C}$, $1 \leq k_1 \leq k_2 \leq K$ détermine fonctionnellement un autre ensemble (disjoint) de concepts $Y = \{Y_{k_3}, \ldots, Y_{k_4}\} \subseteq \mathcal{C}$, $1 \leq k_3 \leq k_4 \leq K$ si et seulement si $\forall n_1, n_2 \in [1, N]$:

$$(\forall X_k \in X, x_{k,n_1} = x_{k,n_2}) \Rightarrow (\forall Y_{k'} \in Y, y_{k',n_1} = y_{k',n_2}).$$

   Ce type spécifique de relations est nécessaire dans notre approche pour formaliser certaines dépendances entre variables. Une dépendance fonctionnelle FD entre X et Y définit une application depuis les valeurs prises par les variables X vers les valeurs prises par les variables Y. Cette fonction sera notée $DetVal_{FD}$:

$$DetVal_{FD} : Range(X_{k_1}) \times \ldots \times Range(X_{k_2}) \rightarrow Range(Y_{k_3}) \times \ldots \times Range(Y_{k_4}).$$

   Deux instances de dépendances fonctionnelles sont requises dans notre approche :

   - une relation *propriété* $\mathcal{P} : \mathcal{C} \rightarrow 2^{|\mathcal{C}|}$ associant, à un concept seul, un ensemble d'autres concepts, qui représente un ensemble de propriétés associées.
     Pour chaque concept associé à ses propriétés, c'est-à-dire $\forall c \in \mathcal{C}$, $\mathcal{P}(c) \neq \emptyset$, nous notons p_c le nombre de propriétés et $\mathcal{P}(c)_i$ le i^e élément de $\mathcal{P}(c)$, $i = 1, \ldots, p_c$.
     Exemple 3.1 *Considérons le concept Couscous de la figure 3.2, une spécialisation de Durum wheat food product. Couscous peut être*

caractérisé par ses propriétés grainsize (de valeur small, medium ou large) et type (de valeur white ou whole-grain), c'est-à-dire :

$$\mathcal{P}(Couscous) = \{Grain\ size,\ Type\}$$

et on a $\mathcal{P}(Couscous)_2 = Type$.
Exemple 3.2 *En chimie, le concept Molecule peut être associé aux propriétés suivantes (et bien d'autres) :*

$$\mathcal{P}(Molecule) = \{Hydrosoluble,\ Molar\ mass\},$$

La fonction $DetVal_{\mathcal{P}}$ sera notée $\mathcal{HP}_c$ (pour $HasProperty$). Elle associe à un élément de $Range(c)$ les valeurs de propriétés spécifiques qu'il prend [2] dans les domaines des concepts de $\mathcal{P}(c)$:

$$\mathcal{HP}_c : Range(c) \rightarrow Range(\mathcal{P}(c)_1) \times \ldots \times Range(\mathcal{P}(c)_{p_c}) \qquad (3.2)$$

Nous notons $\mathcal{HP}_{c\downarrow i} : Range(c) \rightarrow Range(\mathcal{P}(c)_i)$ la restriction de $\mathcal{HP}_c$ à sa i^{eme} propriété, c'est-à-dire $\mathcal{HP}_{c\downarrow i} = \mathcal{HP}_c \cap (Range(c) \times Range(\mathcal{P}(c)_i)$.
Exemple 3.3 *Pour le sous-concept particulier White small-grain couscous* $\in Range(Couscous)$*, on a :*

$$\mathcal{HP}_c(White\ small\text{-}grain\ couscous) = \{small, white\}$$

et $\mathcal{HP}_{c\downarrow 2}(White\ small\text{-}grain\ couscous) = white$.
Exemple 3.4 *Reprenant l'exemple en chimie, nous avons pour NaCl (sel) :*

$$\mathcal{HP}_c(NaCl) = \{Yes, 58.4\}$$

- une relation *définit* $\mathcal{D} : 2^{|\mathcal{C}|} \rightarrow \mathcal{C}$ qui associe, aux valeurs d'un sous-ensemble de concepts, la valeur prise par un autre concept. Des exemples typiques de cette relation sont les équations liant des paramètres d'entrée à un paramètre de sortie.
  Exemple 3.5 *Lorsqu'on caractérise les transferts de gaz à travers un matériau, la perméation (une propriété du matériau) est définie comme le produit de l'épaisseur par la perméance du matériau (cette*

2. Notons qu'un ensemble donné de propriétés ne détermine pas de façon unique un sous-concept satisfaisant ces propriétés, i.e. la relation n'est pas injective.

dernière caractérise le taux de transfert de gaz à travers le matériau, ramené à la différence de pression partielle). Si permeation, permeance et thickness sont trois concepts, alors :

$$\mathcal{D}(\{permeance, thickness\}) = permeation$$

modélise le fait que la valeur de perméation peut être inférée à partir des valeurs de perméance et d'épaisseur.
La fonction $DetVal_{\mathcal{D}}$ sera notée $\mathcal{HD}_C$ (pour $HasDefinition$).
$\forall C \in 2^{|C|}$ tel que $\mathcal{D}(C) \neq \emptyset$, nous définissons la fonction $\mathcal{HD}_C$ telle que :

$$\mathcal{HD}_C : Range(c_1) \times \ldots \times Range(c_{|C|}) \rightarrow Range(\mathcal{D}(C)). \quad (3.3)$$

où c_i et $|C|$ sont respectivement le i^{eme} élément et le nombre d'éléments de C. La fonction $\mathcal{HD}$ définit les valeurs de $\mathcal{D}(C)$, étant données les valeurs des variables déterminantes. Dans le cas de la perméation donné précédemment, le résultat est le produit des deux variables déterminantes.

La figure 3.3 fournit un exemple de trois variables symboliques hiérarchisées : *pH*, *Water* et *Thermosensitivity*, avec leurs sous-ontologies induites par l'ordre $\preceq$. Notons que *pH* est un exemple de variable numérique continue, discrétisée en variable symbolique, et que la sous-ontologie décrivant *Water n'est pas* une taxinomie simple. Nous ferons référence à cette figure et aux variables représentées à plusieurs reprises dans nos exemples à venir.

3.4 Traitement des données utilisant l'ontologie

Lors de l'apprentissage d'un modèle prédictif à partir de données, certaines variables d'entrée et/ou leurs modalités peuvent manquer de pertinence, au moins pour l'expert, même si elles sont statistiquement significatives. En effet, les données expérimentales relatées dans des publications, des rapports, etc., ont généralement été collectées pour des objectifs de recherche très spécialisés et peuvent être inadaptées à une approche d'ingénierie des connaissances. L'objectif des techniques de préparation des données proposées ci-dessous est de remplacer des variables non pertinentes par des

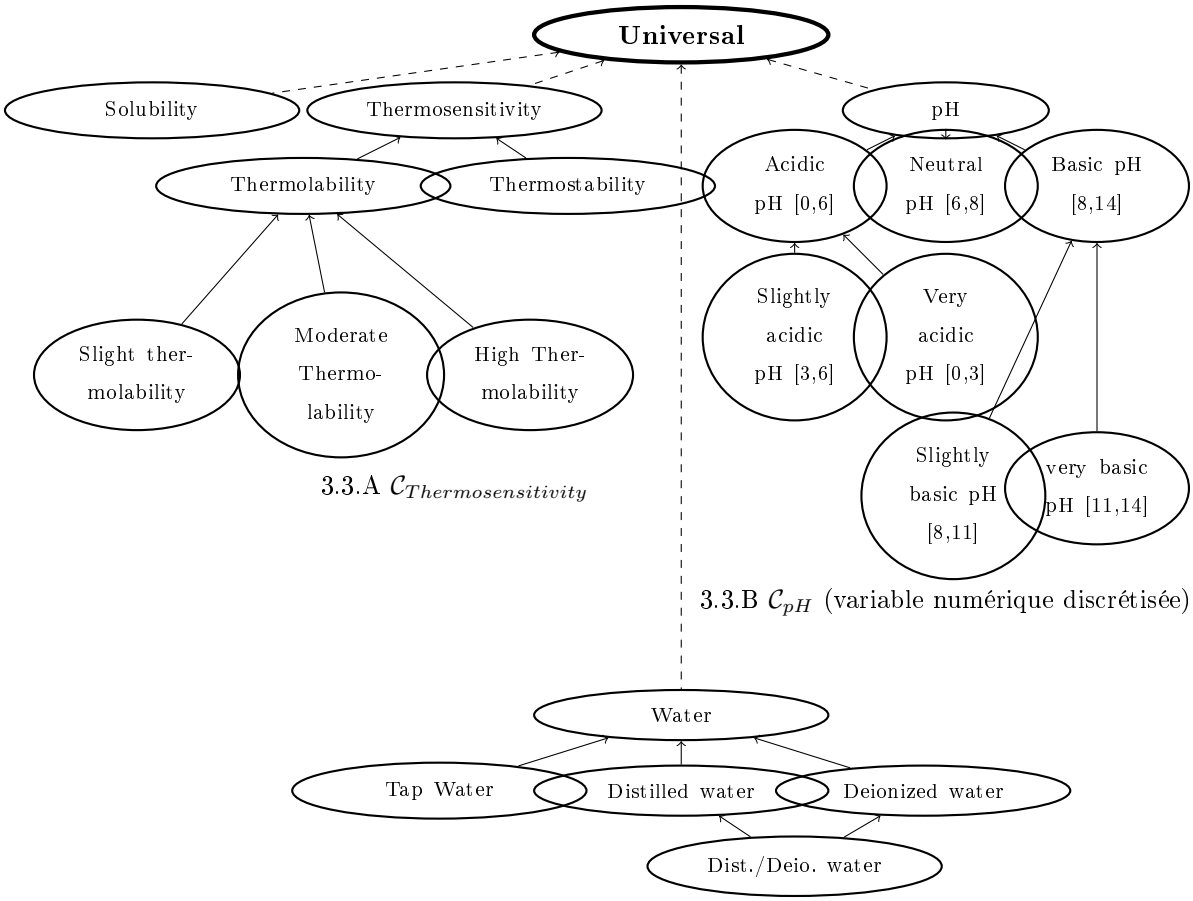

FIGURE 3.3 – Quelques variables et parties de l'ontologie associées, où $A \rightarrow B$ signifie que A est une *sorte de* B

variables significatives pour l'expert. Ce faisant, il est à espérer que leur traitement à venir aboutira à des modèles plus porteurs de sens, auxquels l'expert accordera plus de confiance.

Les techniques présentées ici nécessitent à la fois l'ontologie et les retours de l'expert. De tels retours peuvent être stimulés par une méthode de traitement tierce, telle que les arbres de décision comme ici, des bases de règles floues, etc. Les techniques appropriées à utiliser sur un jeu de données dépendent des retours de l'expert. Nous présentons ici quelques transformations que nous jugeons communes à la plupart des sciences expérimentales, toutefois il peut exister des situations où des relations additionnelles spécifiques seront requises.

3.4.1 Remplacement d'une variable par de nouvelles variables

Ce processus consiste à remplacer une variable donnée par certaines de ses propriétés (plus pertinentes), qui deviennent de nouvelles variables. Soit X_k une variable telle que $\mathcal{P}(X_k) \neq \emptyset$. Pour chaque propriété $\mathcal{P}(X_k)_i$, $i \in [1; p_{X_k}]$ (ou un sous-ensemble d'entre elles), nous créons une nouvelle variable X_{K+i} telle que :

$$\forall n \in [1; N] \quad x_{K+i,n} = \mathcal{HP}_{X_k}(x_{k,n})_{\downarrow i} \tag{3.4}$$

où $\mathcal{HP}_{X_k}(x_{k,n})_{\downarrow i}$ est la projection de $\mathcal{HP}_{X_k}(x_{k,n})$ sur $Range(\mathcal{P}(X_k)_i)$ et $\mathcal{P}(X_k)_i$ le i^{eme} élément de $\mathcal{P}(X_k)$.

Exemple 3.6 *Soit X_k = Couscous la variable (non pertinente) à remplacer et $\mathcal{P}(Couscous)$ = {Grain size, Type} ses propriétés retenues comme pertinentes. Les deux nouvelles variables créées à partir de Couscous sont X_{K+1} = Grain size et X_{K+2} = Type.*

Supposons maintenant que pour la i^{eme} expérience, $x_{k,i}$ = White small-grain couscous. Alors les deux nouvelles valeurs pour la i^{eme} expérience sont $x_{K+1,i} = \mathcal{HP}_{X_k\downarrow 1}(x_{k,i})$ = small et $x_{K+2,i} = \mathcal{HP}_{X_k\downarrow 2}(x_{k,i})$ = white. La variable initiale X_k = Couscous est supprimée.

3.4.2 Regroupement de modalités d'une variable sur la base de propriétés communes

Il peut être utile dans certains cas de considérer des sous-ensembles de modalités présentant une propriété particulière plutôt que les modalités elles-mêmes. Cela équivaut à considérer des parties de l'ensemble des modalités.

Soit X_k une variable telle que $\mathcal{P}(X_k) \neq \emptyset$ et soit $i \in [1; p_{X_k}]$. Nous remplaçons X_k par X'_k tel que, pour $n \in [1; N]$:

$z_n = \mathcal{HP}_{X_k}(x_{k,n})_{\downarrow i}, z_n \in Range(\mathcal{P}(X_k)_i)$ et $x'_{k,n} = \mathcal{HP}_{X_k\downarrow i}^{-1}(z_n)$.

La première équation exprime que nous prenons d'abord z_n, la valeur de la i^{eme} propriété associée à $x_{k,n}$. La seconde équation exprime la recherche de tous les antécédents, c'est-à-dire de tous les $x_{k,l}$ ($l \in [1; N]$) dont la i^{eme} propriété prend la valeur z_n, ce qui inclut $x_{k,n}$ mais peut aussi inclure d'autres valeurs.

Exemple 3.7 *Soit $X_k = Water$ et $pH \in \mathcal{P}(Water)$. Supposons que nous voulions garder la trace des types d'eau utilisés dans les expériences mais qu'il soit souhaitable de les grouper par pH. Nous avons $\mathcal{HP}_{Water}(Tap\ water)_{\downarrow pH} = Basic\ pH$ et $\mathcal{HP}_{Water}(c)_{\downarrow pH} = Neutral\ pH$ pour tout autre $c \in \mathcal{C}_{Water}$. La nouvelle variable X'_k a donc les deux modalités suivantes : $\{Tap\ Water\}$ et $\{Deionized\ water, Distilled\ water, Distilled\ deionized\ water\}$. La seconde modalité étant multi-valuée, elle peut être remplacée par un nouveau concept Ion-poor water dans $\mathcal{C}$, ajouté comme sous-concept de Water et comme sur-concept de Distilled water et Deionized water (voir figure 3.3).*

3.4.3 Fusion de variables pour créer une nouvelle variable

Fusionner plusieurs variables en une nouvelle variable peut être pertinent, les valeurs de la nouvelle variable étant déterminées par les celles des variables fusionnées. L'intérêt est à la fois de faciliter l'interprétation (puisqu'il y a moins de variables à prendre en compte) et d'éviter de considérer comme significative une variable isolée qui n'est significative (du moins du point de vue de l'expert) que conjointement avec d'autres variables.

Soit $C = \{X_{k_1}, \ldots, X_{k_{|C|}}\} \in 2^{\mathcal{X}}$ tel que $\mathcal{D}(C) \neq \emptyset$. Nous définissons alors une nouvelle variable : $\mathcal{X}_{K+1} = \mathcal{D}(\{X_{k_1}, \ldots, X_{k_{|C|}}\})$ telle que : $\forall n \in [1; N]$ $x_{K+1,n} = \mathcal{HD}_C(\{x_{k_1,n}, \ldots, x_{k_{|C|},n}\})$.

Exemple 3.8 *Lors de la cuisson des pâtes, les experts en technologies des céréales font la distinction entre les produits en sous-cuisson, en sur-cuisson ou en cuisson optimale (Under-cooked, Over-cooked, Optimally cooked). Toutefois, ces états dépendent du type de pâtes et du temps de cuisson, et ce sont généralement ces variables qui apparaissent dans les expériences. D'où l'intérêt de remplacer Cooking time et Pasta type par une nouvelle variable Cooking type. Par exemple,* $\mathcal{HD}_C(\{18min, Short\}) = Over\text{-}cooked$*, permet de remplacer Cooking time=18 et Pasta type=Short par Cooking type=Over-cooked, dans toutes les expériences correspondantes.*

3.5 Approche interactive : principes et évaluation

Dans cette partie, nous présentons d'abord les principes de l'approche interactive adoptée. Puis nous indiquons les moyens d'évaluation de l'approche et de ses résultats.

3.5.1 Principes

Nous supposons que nous partons d'une ontologie du domaine initiale $\Omega_0 = \{\mathcal{C}_0, \mathcal{R}_0\}$, qui peut être obtenue par des méthodes semi-automatiques (voir chapitre 2, par élicitation de connaissances expertes ou déjà disponible. Nous supposons également qu'un ensemble de données d'apprentissage initial $\mathbb{D}_0$ est disponible, dont les variables coïncident avec les concepts de l'ontologie.

Méthodes d'apprentissage et connaissances ontologiques sont combinées dans un processus interactif et itératif représenté sur la figure 3.1. L'itération i peut être résumée comme suit :

1. induction du modèle $\mathbb{M}_i$ à partir du jeu de données $\mathbb{D}_i$ (en commençant par $\mathbb{D}_0$) ;
2. évaluation numérique de la qualité du modèle $\mathbb{M}_i$ et discussion de son interprétation avec les experts du domaine ;
3. si le modèle est considéré comme satisfaisant par les experts, fin du processus, sinon, élicitation des transformations à effectuer sur les variables, auprès des experts. Ajout des concepts et des relations nouvel-

lement identifiés à l'ontologie Ω_i (en commençant par Ω_0), pour obtenir Ω_{i+1} ;

4. à partir d'Ω_{i+1} et des indications des experts, transformation des données $\mathbb{D}_i$ (en utilisant les méthodes de la partie 3.4) pour obtenir $\mathbb{D}_{i+1}$;
5. incrémentation de i et retour à l'étape 1.

3.5.2 Evaluation

Deux types d'évaluation de la méthode proposée sont possibles :

• *l'évaluation subjective :* l'évaluation par les experts du domaine qui indiquent leur confiance dans les résultats et les éventuelles inconsistances détectées dans le modèle ;

• *l'évaluation objective :* l'évaluation numérique automatique, où les résultats et la stabilité du modèle sont évalués par des indicateurs numériques :

– Le critère classique pour les arbres de classification est le taux d'erreur, $Ec = \frac{MC}{N}$, où MC est le nombre de données mal classées et N la taille du jeu de données, calculé par validation croisée ou sur l'ensemble du jeu de données.
– La complexité de l'arbre est : $Nrules + Nnodes/Nrules$, où $Nrules$ est le nombre de noeuds terminaux (feuilles), ce qui équivaut au nombre de règles, et $Nnodes$ est le nombre total de noeuds dans l'arbre.

Lors de l'évaluation, il est également important de prendre en compte la fiabilité des données : sources, équipement expérimental, protocole ...

3.6 Etude de cas : application à la prédiction de la qualité alimentaire

L'étude de cas concerne les procédés de transformation des céréales et en particulier du blé dur. Elle s'intéresse à l'impact de ces procédés sur la qualité nutritionnelle des produits finis.

Les systèmes existants proposés en sciences de l'aliment et notamment en transformation des céréales pour aider la prédiction (Young, 2007), n'exploitent pas la double source d'information constituée par les données expérimentales et les connaissances expertes. Ils ne proposent pas non plus de

solution en l'absence d'un modèle prédéterminé (mathématique ou expert).

3.6.1 Contexte et description de l'étude de cas

Les information concernant l'impact de chaque opération unitaire du procédé de transformation sur chaque type de propriété du produit, se trouvent sous la forme d'un ensemble de données. Les variables d'entrée sont les paramètres de l'opération unitaire. La variable de sortie est l'impact de l'opération sur une propriété (par exemple la variation de la teneur en vitamines). Nous étudions ici le cas de l'opération unitaire *Cuisson à l'eau* et de la propriété *Teneur en vitamines*. Ce cas concerne 150 données expérimentales et implique 60 concepts de l'ontologie. Le tableau 3.2 présente quelques valeurs des variables d'entrée et de la variable de sortie. L'ontologie a été créée avec CoGUI (*http ://www.lirmm.fr/cogui/*). La transformation des données et les arbres de décision ont été obtenus avec le logiciel R (R Development Core Team, 2009) (utilisation du package *R-WEKA* et environ 2000 lignes de code développé).

Id	Vitamin	Cooking temp. (C)	Cooking time (min)	Water	Vitamin loss (%)
1	B6	100	13	NA	-52
2	B2	100	12	Tap	-53
3	B1	98	15	Distilled	-47
4	B2	90	10	NA	-18
5	B1	100	NA	Dist./Deio.	-41

TABLE 3.2 – Une partie du jeu de données d'entraînement

3.6.2 Application de l'approche à l'étude de cas

La variable de sortie est la *Variation de la teneur en vitamines* durant l'opération unitaire. Il s'agit d'une variable continue, discrétisée en quatre classes ordonnées *Low loss, Average loss, High loss, Very high loss*.

Tous les arbres sont construits avec un nombre minimum d'instances par feuille égal à 6, puis élagués. Les graphiques sont à interpréter de la façon suivante :

1. Chaque noeud de test est étiqueté par la variable de segmentation ;

2. pour chaque feuille, le nombre d'observations mal classées est spécifié.

Notre approche se conforme au processus itératif présenté dans la partie 3.5 et est illustrée par quatre itérations.

Itération 1 : état initial

La figure 3.4 montre l'arbre obtenu à partir du jeu de données brutes ($\mathbb{D}_0$). Comme mentionné dans la partie 3.2.3, la complexité des arbres de décision C4.5 augmente avec le nombre de modalités, comme c'est le cas pour la variable *Kind of water*. Le but de notre approche est aussi de réduire cette complexité en identifiant les propriétés sous-jacentes pertinentes cachées derrière ces modalités.

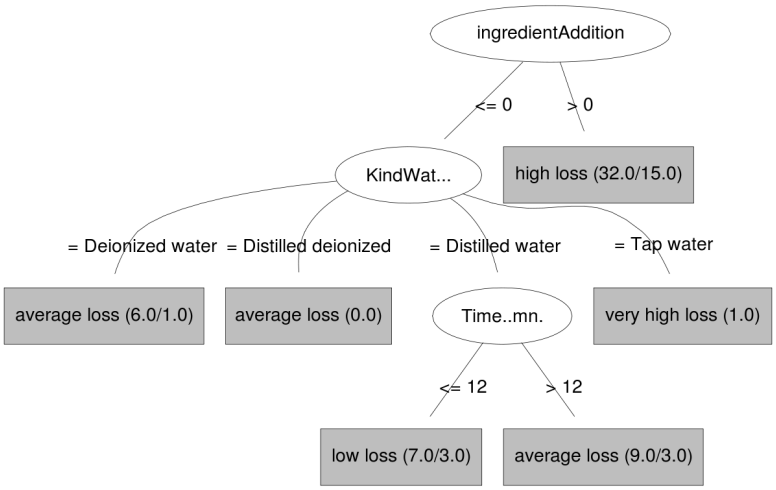

FIGURE 3.4 – Arbres de décision sur les données brutes

L'examen de l'arbre par les experts du domaine a abouti aux remarques et ajustements suivants. Tout d'abord, la variable la plus discriminante est *Ingredient Addition*. En effet, elle correspond à l'ajout de vitamines pour compenser les pertes dues à l'opération de cuisson. Les experts suggèrent d'*enrichir l'ontologie* en caractérisant les vitamines par leurs propriétés. Les éléments suivants ont été ajoutés à l'ontologie (pour obtenir Ω_1) et les données

ont été transformées pour obtenir $\mathbb{D}_1$.

$$\mathcal{P}(Vitamin)=\{Solubility,Thermosensitivity,Photosensitivity,...\}$$

$$Range(Photosensitivity)=\{Photolabile,Photostabile\}$$

$$\mathcal{HP}_{Vitamin}(VitaminA)=\{Liposoluble,Thermolabile,Photostabile\}$$

Itération 2 : introduction de connaissances sur les propriétés des vitamines

Le modèle $\mathbb{M}_1$ est un nouvel arbre illustré sur la figure 3.5. Les variables *Kind of water* et *Cooking time* apparaissent dans cet arbre. La discussion

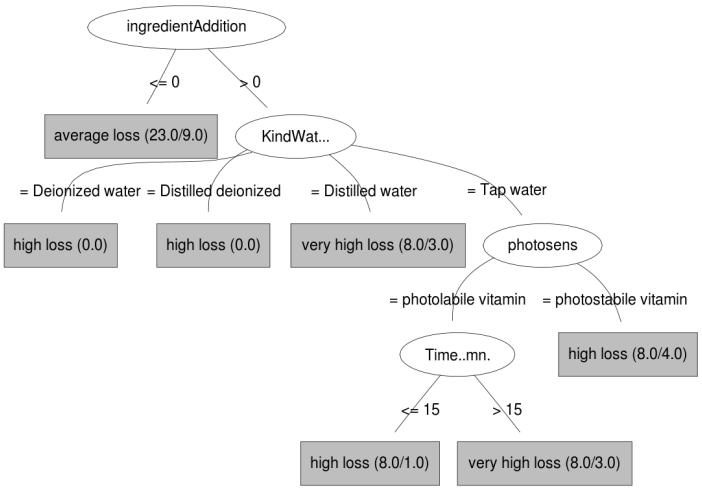

FIGURE 3.5 – Arbres de décision sur les données avec les propriétés des vitamines

avec les experts met alors en évidence que la variable *Cooking time* n'est pertinente que si elle est considérée conjointement avec la variable *Pasta type*. Les experts indiquent également que l'eau peut être mieux décrite par son *pH* et sa dureté (*Hardness*). Dans les expériences disponibles, le pH et la dureté de l'eau n'étaient pas mesurés. Ils peuvent cependant être retrouvés à partir du type d'eau. Les éléments suivants ont été ajoutés à Ω_1 pour obtenir

Ω_2 et utilisés pour transformer $\mathbb{D}_1$ en $\mathbb{D}_2$ (voir section 3.3) :

$\mathcal{P}(Water)$=$\{pH$,$Hardness\}$, $Range(ph)$=$\{AcidpH$,$NeutralpH$,$BasicpH\}$

$\mathcal{HP}_{water}(Tapwater)$=$\{NeutralpH$,$Hard\}$

$\mathcal{D}(\{Pastatype$,$Cookingtime\})$=$Cookingtype$, $\mathcal{HD}(\{short$,$18min\})$=$Overcooking$

Itération 3 : introduction du type de cuisson et des propriétés de l'eau

La figure 3.6 montre $\mathbb{M}_2$, l'arbre obtenu avec les modifications précédentes. On peut voir sur cet arbre que la variable *Hardness* nouvellement

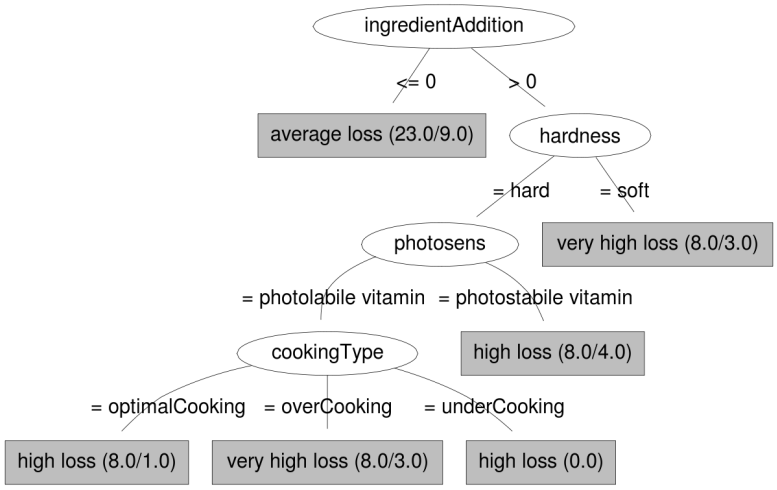

FIGURE 3.6 – Arbre de décision avec le type de cuisson et les propriétés de l'eau

créée est maintenant sélectionnée à la deuxième segmentation. La discussion avec les experts fait ressortir l'existence d'un lien entre la dureté de l'eau et l'évolution du pH. L'évolution du pH de l'eau dépend à la fois de la température de cuisson et de la dureté de l'eau. Une nouvelle variable est alors créée suivant des règles expertes qui ne sont pas détaillées ici, aboutissant à Ω_3 et $\mathbb{D}_3$.

$\mathcal{D}(\{pH, Temperature\}) = CookingpH$

Itération 4 : introduction du pH de cuisson

La figure 3.7 affiche l'arbre C4.5 final (modèle $\mathbb{M}_3$). Les variables sélec-

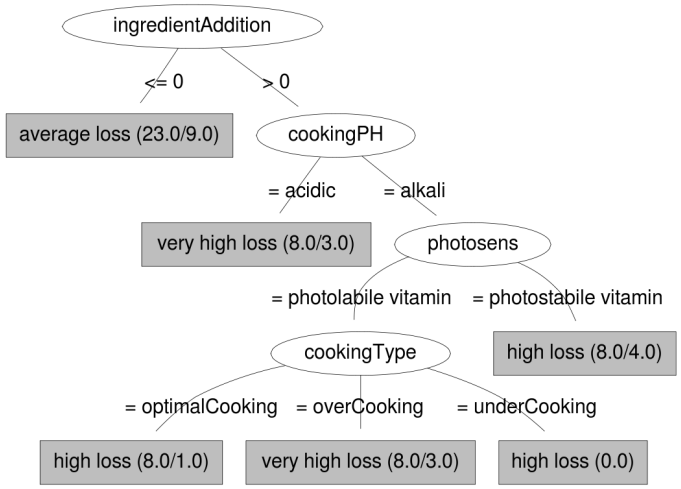

FIGURE 3.7 – Arbre de décision à l'état final

tionnées par l'algorithme d'apprentissage sont maintenant jugées pertinentes. En particulier, certaines variables initiales mesurées et continues, comme *Cooking time*, sont maintenant remplacées par des variables plus significatives, telles que *Cooking type* obtenu suite à l'ajout d'un nouveau concept (*Pasta type*) dans l'ontologie.

Le tableau 3.3 présente l'évolution des critères définis dans la partie 3.5.2.

Itération #	taux MC (%)	Complexité
1	44	7.3
2	48	8.4
3	35	7.5
4	35	7.5

TABLE 3.3 – Evaluation de l'arbre

Bien que le taux d'erreur demeure élevé, essentiellement du fait de la faible quantité de données, il s'améliore aux deux dernières itérations, tandis que la complexité demeure limitée. L'examen de la matrice de confusion

montre ensuite que la plupart des erreurs de prédiction concernent des cas où l'étiquette attribuée est proche de celle attendue, par exemple *High Loss* au lieu de *Very High Loss*.

3.7 Conclusion du chapitre

La formalisation et l'acquisition de nouvelles connaissances expertes, ainsi que la construction de modèles fiables, sont deux aspects essentiels de la recherche en intelligence artificielle dans les sciences expérimentales. Une importance particulière est à attacher à la confiance que les experts du domaine accordent aux modèles appris. Comme dans d'autres domaines (web sémantique, ...), les connaissances apprises à partir des données et ontologiques peuvent s'enrichir mutuellement.

Dans ce chapitre, nous avons présenté une approche collaborative et itérative, où les connaissances expertes et les opinions concernant les modèles appris ont été intégrés à l'ontologie décrivant les connaissances du domaine. Cette formalisation est ensuite réutilisée pour transformer les données disponibles et en apprendre de nouveaux modèles. Ces nouveaux modèles sont alors de nouveau la source d'opinions expertes complémentaires et ainsi de suite, jusqu'à l'obtention de résultats satisfaisants aux yeux des experts. Cette démarche permet à la fois d'enrichir les connaissances ontologiques et d'accroître la confiance accordée par les experts aux résultats des méthodes d'apprentissage.

La méthode présentée est appliquée à une étude de cas dans le domaine de la transformation des céréales. Cette étude de cas a été réalisée de façon itérative, en collaboration étroite avec les experts du domaine. Elle démontre la valeur ajoutée de la prise en compte des connaissances ontologiques, en améliorant l'interprétabilité et la pertinence des résultats. Elle vise aussi, en présentant les modèles appris aux experts, à faire ressortir des connaissances ontologiques potentiellement utiles dans d'autres applications.

Ce travail est une étape vers l'objectif difficile de construire des méthodes semi-automatiques. Une de ses perspectives est la recherche d'une plus grande automatisation de l'ensemble du processus.

Chapitre 4

Argumentation pour l'aide à la décision

La décision argumentée est un thème encore peu abordé dans la littérature internationale. Elle fait suite aux travaux sur l'argumentation qui est un domaine de recherche relativement récent en intelligence artificielle. Ce chapitre présente un modèle général pour établir des recommandations s'appuyant sur l'argumentation, en étendant le système d'argumentation fondateur de Dung. Cette approche est appliquée à l'analyse des arguments concernant la qualité alimentaire dans une politique de santé publique. Les produits céréaliers, et le pain en particulier, sont utilisés par les décideurs comme levier pour combattre des maladies telles que l'obésité ou le diabète. Le modèle proposé fournit des recommandations s'appuyant sur l'argumentation des parties prenantes et visant des publics particuliers.

Ces travaux ont été réalisés dans le cadre de la thèse de Jean-Rémi Bourguet (Bourguet, 2010) et en collaboration avec Leila Amgoud, Marie-Laure Mugnier, Jérôme Fortin et Joël Abécassis. Différents aspects étudiés, dont une partie est présentée ici, ont été publiés notamment dans (Bourguet et collab., 2009, 2010; Fortin et collab., 2011; Bourguet et collab., 2013).

4.1 Problématique

La définition de la qualité d'un produit alimentaire repose sur de nombreux critères, nutritionnels, organoleptiques, d'usage et sanitaires, voire plus

récemment environnementaux et économiques. Ces différents critères ne sont pas toujours compatibles. Ainsi, la consommation de produits céréaliers complets, si bénéfique soit-elle d'un point de vue nutritionnel par son apport en micro-nutriments et en fibres, pose la question du risque de contamination, par exemple par les phytosanitaires. Faut-il choisir le bio ou se fier aux réglementations, peut-on tout sacrifier à la sécurité et à la santé au risque de sous-estimer le plaisir gustatif ? Le dilemme se pose pour les consommateurs, mais concerne également les acteurs des filières et les décideurs.

Un compromis entre qualité nutritionnelle, organoleptique et sanitaire s'est construit de manière empirique au sein des filières, avec la maîtrise progressive des procédés de transformation. L'émergence de nouvelles préoccupations et de nouvelles demandes implique aujourd'hui le déplacement de ce compromis vers un nouvel équilibre, faisant en particulier une plus large place aux aspects nutritionnels. Ainsi les politiques de santé publique, dont fait partie le PNNS (Programme National Nutrition Santé) lancé en 2001, tentent de faire face à certaines maladies en augmentation dans nos sociétés occidentales (maladies cardiovasculaires, cancers, obésité, ...). Les consommateurs se sensibilisent à ces nouvelles problématiques et les industriels ont besoin d'outils pour répondre aux demandes émergentes par l'adaptation, l'innovation, l'optimisation des procédés de transformation au sein des filières.

L'importance accordée à ces différents critères varie en fonction des acteurs considérés. Ainsi les experts peuvent estimer un niveau de risque lié à un contaminant comme tout à fait acceptable au vu du bénéfice apporté (par exemple pour un produit phytosanitaire) ou comparé au coût de précautions supplémentaires peu efficaces (par exemple pour une mycotoxine), tandis que le consommateur n'acceptera pas l'existence d'un risque alimentaire même minime. Le questionnement scientifique sous-jacent est le suivant :

- Quel modèle de représentation est adapté à la prise en compte de ces points de vue contradictoires ?
- Comment prendre en compte les priorités des différents acteurs concernés et l'importance relative qu'ils accordent aux critères considérés ?
- Peut-on imaginer des scénarios différents représentatifs des préoccupations des différents acteurs concernés ?
- Comment résoudre les conflits posés pour parvenir à un compromis au sein d'un système d'aide à la décision ?

Ce chapitre présente une approche qui permet, d'une part, la formalisa-

tion des connaissances disponibles en tant qu'éléments de prise de décision, incluant l'expertise "implicite" et pas uniquement les données analytiques plus classiquement utilisées ; d'autre part, l'aide à la décision par l'utilisation des priorités adaptées en fonction du public visé. L'approche s'appuie sur les systèmes d'argumentation, modèles de raisonnement fondés sur la construction et l'évaluation des arguments en interaction, éventuellement conflictuels.

Parmi les approches existantes, soulignons que les méthodes fondées sur la comparaison de fonctions de risque, de type numérique, ne sont pas applicables ici, dans la mesure où le travail s'appuie en grande partie sur la formalisation de connaissances expertes implicites, de nature symbolique. La décision multicritère classique, fondée sur l'évaluation de plusieurs options possibles en s'appuyant sur un ensemble de critères, est également inadaptée car elle ne permet pas la représentation de points de vue contradictoires et de débats. Le lecteur pourra se référer par exemple à (Figueira et collab., 2005; Bouyssou et collab., 2009) pour des précisions et des synthèses sur les méthodes de décision existantes.

A l'heure actuelle, les démarches les plus proches d'une prise de décision en présence de points de vue contradictoires sont les travaux traitant des systèmes argumentatifs (Besnard et Hunter, 2008; Rahwan et Simari, 2009). Le raisonnement argumentatif a été étudié pour sa capacité à analyser des situations où les informations sont incohérentes parce qu'elles proviennent de différentes sources ou correspondent à différents points de vue potentiellement divergents. Il apparaît également avoir un rôle important dans les tâches de décision où avantages et inconvénients doivent être appréciés à partir des connaissances disponibles. Argumentation et décision ont jusqu'ici été abordés séparément, avec des objectifs différents. Les mécanismes d'argumentation permettant d'*introduire des éléments d'explication* dans la prise de décision sont très peu abordés dans la littérature (Amgoud et Prade, 2009).

Ce chapitre concerne deux aspects. D'une part, il analyse une étude de cas concernant l'évaluation risque/bénéfice dans la filière boulangère, suite aux recommandations du PNNS en faveur de produits céréaliers plus complets. En effet, cette recommandation doit faire face à des points de vue différents et à de fortes réserves de la part des professionnels de la filière blé. Cette étude de cas s'appuie sur l'analyse de nombreuses sources d'information : articles scientifiques, documents techniques, entretiens, conférences et débats. D'autre part, ce chapitre décrit un modèle d'argumentation générique permettant la représentation et l'évaluation de cette étude de cas mais éga-

lement applicable à d'autres domaines. Le chapitre présente successivement la méthodologie adoptée (partie 4.2) et les résultats obtenus (partie 4.3).

4.2 Méthodologie

Cette partie décrit les étapes de la méthodologie : l'identification des sources d'information disponibles (partie 4.2.1), la modélisation des arguments (partie 4.2.2) et les principes des systèmes d'argumentation (partie 4.2.3).

4.2.1 Identification et analyse des sources d'information

Nous nous sommes appuyés sur différentes sources d'information complémentaires comprenant, des plus formelles aux moins formelles :

1. des articles scientifiques évalués par des pairs ;
2. des rapports techniques ou des informations publiées sur les sites Web ;
3. des conférences et réunions scientifiques autour de projets de recherche ;
4. des connaissances expertes obtenues au moyen d'entretiens.

Les articles scientifiques que nous avons analysés incluent (Bourre et collab., 2008; Slavin et Green, 2007; Dubuisson-Quellier, 2006; Ginon et collab., 2009; Layat, 2011). (Bourre et collab., 2008) compare les différents types de farines d'un point de vue nutritionnel. (Slavin et Green, 2007) étudie le lien entre les fibres et la satiété. (Dubuisson-Quellier, 2006; Ginon et collab., 2009) traitent des comportements de consommation et du consentement à payer des consommateurs, en particulier concernant la baguette française lorsque l'information concernant le niveau de fibres est fournie, sur la base expérimentale et statistique d'études de panels de consommateurs. (Layat, 2011) fournit une synthèse sur les aspects nutritionnels de la consommation de pain et fait le lien avec les aspects technologiques.

Nous avons également examiné des rapports techniques disponibles sur les sites web officiels concernant la politique de santé publique du PNNS (Programme National Nutrition-Santé) (PNNS (documents statutaires), 2010; PNNS (site web), 2010), le projet européen Healthgrain sur l'amélioration de la nutrition et de la santé à travers les grains de céréales (Dean et collab., 2007; HEALTHGRAIN, 2009), assisté à des projets et colloques français

au sujet des mesures sanitaires, nutritionnelles, technologiques et organoleptiques des pains (DINABIO, 2008; CADINNO, 2008; AQUANUP, 2009; FCN, 2009).

Enfin, plusieurs entretiens ont été menés pour recueillir les connaissances expertes du domaine, en particulier celles de spécialistes de notre laboratoire, Joël Abécassis et Xavier Rouau, ainsi que d'experts exterieurs, Gérard Brochoire (INBP : Institut National de la Boulangerie Pâtisserie) et Hubert Chiron (INRA de Nantes).

4.2.2 Modélisation des informations disponibles en arguments structurés

A partir des sources d'information présentées ci-dessus, le travail de modélisation est une tâche itérative où les arguments sont d'abord recueillis en provenance des diverses parties prenantes, formalisés, puis validés par des experts de différents domaines.

Les motivations du PNNS sont d'abord considérées comme des "raisons" qui justifient des arguments. La première étape de modélisation de l'argument est donc l'extraction d'une raison (notée `Raison`).

Etude de cas. Extraction de la raison *"Considérant la nutrition comme un levier déterminant pour la santé, une priorité nutritionnelle peut être d'augmenter la consommation quotidienne de glucides complexes par une hausse de la teneur en fibres dans les aliments." De cet énoncé issu de PNNS (documents statutaires) (2010); PNNS (site web) (2010), la raison générale du tableau 4.1 peut être extraite.*

Raison
"La hausse de la teneur en fibres dans la diète est pertinente."

TABLE 4.1 – Raison

Cette raison générale peut être raffinée en des raisons plus spécifiques qui soutiennent directement des actions. Par conséquent, dans cette étude de cas, un argument est considéré comme un motif (une raison) soutenant une décision, une recommandation ou plus généralement une action (notée `Action`).

Etude de cas suite. Action soutenue *"Le pain est sélectionné dans ce programme comme une source nutritionnelle de consommation quotidienne de fibres. L'augmentation du rendement en farine (↗ R.F) se traduit par une plus forte teneur en fibres dans la farine et par conséquent dans le pain. Pour ces raisons, le PNNS envisage la possibilité d'une évolution de la législation du pain de consommation courante visant à augmenter la teneur en fibres dans le pain." De cet énoncé issu de PNNS (documents statutaires) (2010); PNNS (site web) (2010), une raison spécifique supporte une action comme indiqué dans le tableau 4.2.*

	Raison	Action
1	"L'augmentation du rendement en farine permet d'augmenter la teneur en fibres."	↗ R.F

TABLE 4.2 – Une raison supporte une action

4.2.3 Modèles d'argumentation existants

L'argumentation est un modèle de raisonnement basé sur la construction et l'évaluation des arguments en interaction. Ce modèle a notamment pu être étudié dans le cadre du raisonnement non monotone (Dung, 1995), en prise de décision (Bonet et Geffner, 1996; Fox et Das, 2000) ainsi que pour modéliser différents types de dialogues incluant notamment la négotiation (Kraus et collab., 1998; Sycara, 1990). Une grande partie des modèles développés pour les applications se sont donc appuyés sur le cadre de travail proposés par (Dung, 1995).

Dans la lignée des travaux existants (Amgoud et Prade, 2009) et pour appréhender les schémas de décision du monde réel, nous adoptons un modèle de décision argumentée dans lequel un argument fournit une raison pour soutenir une recommandation ou pour réaliser une action. Nous y intégrons un ensemble de toutes les options possibles, considérées comme mutuellement exclusives et soutenues par des arguments qui font également partie du modèle. Ce modèle est défini comme suit :

Définition 4.1 *Un* système de décision argumentée `DF` *est un couple* $\langle \mathcal{A}, \mathcal{D} \rangle$ *où :*

- $\mathcal{A}$ *est un ensemble d'arguments ;*

- $\mathcal{D}$ *est un ensemble d'actions, supposées mutuellement exclusives ;*
- `action` $: \mathcal{A} \rightarrow \mathcal{D}$ *est une fonction retournant l'action soutenue par un argument.*

Classiquement, un processus d'argumentation suit trois étapes principales : 1) la construction d'arguments et de contre-arguments, 2) l'évaluation de l'acceptabilité des arguments, 3) la définition de conclusions justifiées.

Dans le cadre de Dung, pour réaliser la première étape, une relation binaire appelée "attaque" est définie sur l'ensemble $\mathcal{A}$, reflétant les conflits entre arguments. Un système d'argumentation à la Dung `AF` peut être construit à partir de `DF`, nous ramenant ainsi à un cadre classique, dans lequel les arguments sont en conflit lorsque les actions qu'ils soutiennent sont disctinctes :

Définition 4.2 *A partir d'un* système de décision argumentée $\langle \mathcal{A}, \mathcal{D} \rangle$*, un système d'argumentation* `AF` $= \langle \mathcal{A}, \mathcal{R} \rangle$ *est construit de la façon suivante :*
- $\mathcal{A}$ *est le même ensemble d'arguments ;*
- $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ *est une relation d'*attaque *telle que* $(\alpha, \beta) \in \mathcal{R}$ *si* `action`$(\alpha) \neq$ `action`(β).

	Raison	Action
1	"L'augmentation du rendement en farine permet d'augmenter la teneur en fibres."	$\nearrow$ R.F
2	"L'augmentation du rendement en farine engendre des bénéfices économiques."	$\nearrow$ R.F
3	"La diminution du rendement en farine rehausse les attributs sensoriels du pain."	$\searrow$ R.F
4	"La diminution du rendement en farine provoque un gain sanitaire."	$\searrow$ R.F

TABLE 4.3 – Arguments soutenant des actions mutuellement exclusives

Etude de cas suite. Représentation du système *Les arguments du tableau 4.3 décrivent différentes raisons soutenant des modifications dans la législation du pain de consommation courante (hausse ou baisse du rendement en farine, notés $\nearrow$ R.F et $\searrow$ R.F). La figure 4.1 montre le graphe d'attaque, représentant ces arguments et la relation d'attaque entre eux.*

Les sémantiques d'acceptabilité de Dung permettent d'identifier, parmi les arguments en conflit, ceux qui seront retenus pour déterminer les décisions acceptables. Une sémantique d'acceptabilité permet de définir des ensembles d'arguments (appelés extensions) satisfaisant un critère de consistance. Les

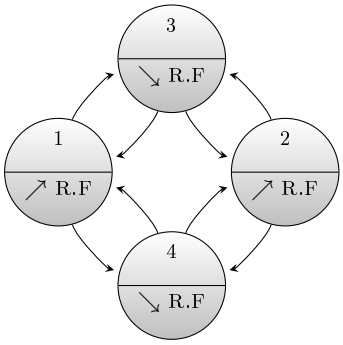

FIGURE 4.1 – Graphe d'attaque associé

principales notions introduites par Dung sont rappelées dans la définition suivante. Notons que d'autres sémantiques ont été définies dans la littérature (voir par exemple Baroni et collab. (2005)).

Définition 4.3 *Soit* $\mathtt{AF} = \langle \mathcal{A}, \mathcal{R} \rangle$ *un système d'argumentation et soit* $\mathcal{B} \subseteq \mathcal{A}$.

- $\mathcal{B}$ *est sans-conflit s'il n'existe pas* $\alpha, \beta \in \mathcal{B}$ *tels que* $(\alpha, \beta) \in \mathcal{R}$.
- $\mathcal{B}$ *défend* α *si pour tout* $\beta \in \mathcal{A}$, *si* $(\beta, \alpha) \in \mathcal{R}$, *alors il existe* $\gamma \in \mathcal{B}$ *tel que* $(\gamma, \beta) \in \mathcal{R}$.
- $\mathcal{B}$ *est une ensemble admissible si* $\mathcal{B}$ *est sans-conflit et défend tous ses éléments.*
- $\mathcal{B}$ *est une extension préférée si* $\mathcal{B}$ *est un ensemble admissible maximal (au sens de l'inclusion ensembliste).*

Il peut exister plusieurs ensembles admissibles. Ces ensembles sont toujours inclus dans au moins une extension préférée, qui peut être utilisée pour la prise de décision. Les décisions recommandées peuvent être obtenues comme étant les actions soutenues par les arguments des extensions préférées. A partir d'un système de décision argumentée, nous définissons un résultat décisionnel noté $\mathtt{result}(\mathtt{DF})$, qui retourne un ensemble de décisions fondées.

Définition 4.4 *Soient* $\mathtt{DF} = \langle \mathcal{A}, \mathcal{D} \rangle$ *un système de décision argumentée et* AF *le système d'argumentation associé. Pour tout* $d \in \mathcal{D}$, $d \in \mathtt{result}(\mathtt{DF})$ *s'il existe une extension préférée* $\mathcal{B}$ *de* AF, *avec* $\alpha \in \mathcal{B}$ *et* $\alpha \in \mathtt{action}^{-1}(d)$.

Etude de cas suite. Décision soutenue *Dans la figure 4.1, chaque argument se défend lui-même et les ensembles d'arguments soutenant la même option sont sans-conflit.*

- *Il y a deux extensions préférées* $\{1, 2\}$ *et* $\{3, 4\}$.
- *Le résultat décisionnel de* `DF` *est l'ensemble* `result(DF)` $= \{\nearrow R.F, \searrow R.F\}$.

Comme le montre l'exemple précédent, on peut obtenir plusieurs options sans qu'il soit possible de n'en retenir qu'une. Des extensions au cadre de Dung intégrant des préférences ont été proposées dans la littérature, notamment celles où les arguments sont supposés avoir différentes forces (Amgoud et Cayrol, 2002; Amgoud et collab., 2000) ou différentes valeurs prioritaires (Bench-Capon, 2003; Kaci et van der Torre, 2008). Le modèle peut donc être affiné par des préférences traduisant les priorités entre arguments. Les arguments eux-mêmes peuvent être affinés par des informations contextuelles propres à l'argumentation (parties prenantes, préoccupations, etc.). Enfin, les actions ne sont pas nécessairement mutuellement exclusives : certaines actions peuvent être plus spécialisées que d'autres. En résumé, le modèle proposé devrait prendre en considération :

- les parties prenantes et les préoccupations (notées `P.Pren` et `Preoccup`) ;
- les buts (notés `But`) promus par les arguments (par exemple une hausse de composant notée $\nearrow$ ou une baisse de composant notée $\searrow$) ;
- les actions spécialisées notées "Action & Spécialisation" (par exemple $\nearrow R.F$ & $S.$ pour "augmentation du rendement en farine et sans sel").

Cette formalisation affinée est détaillée dans la partie 4.3). Elle peut être illustrée par un argument avancé par le PNNS lors de sa campagne initiale, visant à promouvoir un objectif de type nutritionnel.

	P. Pren.	Raison	Action	Préoccup.(s)	But(s)
0	PNNS	"Réduire la consommation moyenne de chlorure de sodium (sel) est pertinente."	$\nearrow$ R.F & S.	Nutrition	$\searrow$ Sel

TABLE 4.4 – Un argument affiné

Etude de cas suite. Affinement des arguments *Dans le tableau 4.4, un argument du PNNS donne une raison de soutenir un pain sans sel, une action notée "*$\nearrow R.F$ & $S.$*" et considérée comme une spécialisation de l'action*

↗ R.F. Cet argument attaque tout argument soutenant une action strictement plus générale (i.e. ↗ R.F). Cette relation d'attaque n'est pas symétrique : si une action a_1 est plus spécifique qu'une action a_2, alors les arguments soutenant a_1 attaquent les arguments soutenant a_2 mais l'inverse n'est pas vrai. Sur cette base, nous proposons de formaliser les principaux arguments des acteurs de la filière boulangère en réponse aux recommandations du PNNS. Le graphe d'attaque obtenu est représenté dans la figure 4.2.

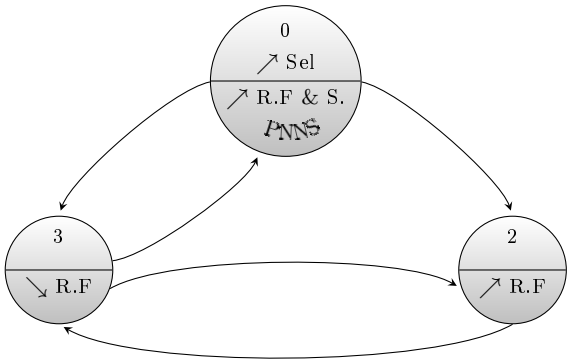

FIGURE 4.2 – Graphe avec attaques non-symétriques

4.3 Résultats

Les arguments, en réponse à la recommandation du PNNS en faveur de produits céréaliers plus complets, sont décrits dans la partie 4.3.2. Un modèle d'argumentation générique pour la représentation, l'évaluation d'arguments et la recommandation d'actions est décrit dans la partie 4.3.3. Les résultats sont résumés dans la partie 4.3.4. Le schéma global de cette approche est présenté dans la partie 4.3.1.

4.3.1 Schéma global

Pour utiliser le modèle proposé, les étapes suivantes doivent être réalisées successivement :

1. Obtenir une représentation de tous les arguments, des buts promus, des actions soutenues, ainsi que des parties prenantes et des préoccupations associées. Les attaques peuvent être définies par la spécialisation d'actions ou par des actions mutuellement exclusives.
2. Définir les audiences, c'est-à-dire les publics ciblés par les arguments, par exemple des segments de consommateurs. Les audiences engendrent des priorités entre buts différentes. Par la suite cette priorisation des buts est utilisée pour caractériser une audience.
3. Générer les relations de préférence entre arguments pour chaque audience, suivant la priorisation des buts.
4. Définir la relation d'attaque entre arguments.
5. Résoudre le système, ce qui conduit à recommander une ou plusieurs actions.

4.3.2 Arguments

Parmi les recommandations nutritionnelles du PNNS figurent les suivantes :

• améliorer le pain courant en termes de qualité nutritionnelle, en développant la consommation des pains fabriqués avec de la farine plus complète, par exemple de type 80 (c'est-à-dire contenant 0.80 g de matière minérale pour 100 g de matière sèche de farine, au lieu de 0.65 g pour 100 g) ;

• établir des chartes d'engagement entre les filières professionnelles (boulangers et meuniers), les sociétés de restauration collective, etc.

Deux options qui s'excluent mutuellement peuvent être caractérisées : passer à l'utilisation de la farine de type 80 (option notée $\curvearrowright T80$) ou conserver celle de type 65 (option notée $\circlearrowleft T65$) pour le pain de consommation courante. Dans le tableau 4.5, sont listés les arguments des décideurs engagés dans la politique de santé publique (`P.Pren` = PNNS), lesquels ont diverses préoccupations (`Preoccup` = Nutrition, Economie), et tentent de promouvoir plusieurs buts (`But` = $\nearrow$ Fibres, $\nearrow$ Oligo-éléments, $\searrow$ Coûts) et soutiennent une option (`Action` = $\curvearrowright T80$).

Ces arguments sont confrontés aux points de vue des autres parties prenantes concernées par la transformation du blé. Notons en premier lieu, les boulangers et les meuniers qui sont inquiets au sujet d'éventuels impacts

	P. Pren.	Raison	Option	Préoccup.(s)	But(s)
1	PNNS	"L'utilisation d'une farine T80 à la place d'une farine T65 pour la panification est pertinente."	$\curvearrowright$ T80	Nutrition	$\nearrow$ Fibres $\nearrow$ Oligo-éléments (o.e.)
2	PNNS	"La T80 réduit les coûts de fabrication en raison d'un meilleur rendement."	$\curvearrowright$ T80	Economie	$\searrow$ Coûts
3	PNNS	"Une diète riche en fibres réduit les coûts de santé publique."	$\curvearrowright$ T80	Economie	$\searrow$ Coûts

TABLE 4.5 – L'argumentation du PNNS

quant à leur cœur d'activité. La meunerie française fait pression pour un réexamen de ces recommandations. Un rapport scientifique d'enquête sur les impacts nutritionnels de la farine de type 80 par rapport aux autres farines du marché a été utilisé pour répondre aux décideurs du PNNS. Dans le tableau 4.6, sont listés les arguments des professionnels de la meunerie (`P.Pren` = Meuniers), au sujet des mêmes préoccupations que le PNNS, et tentant de promouvoir d'autres buts (`But` = $\nearrow$ Offre segmentée, $\nearrow$ Technicité) et soutenant l'option conservatrice (`Action` = $\circlearrowleft T65$) voire une remise en cause de la pertinence de la recommandation, à l'aide par exemple d'un autre indicateur plus représentatif que ne l'est la teneur en cendres (`Action` = $\curvearrowright$ I., où I. signifie Indicateur), qui peut donc être vu comme conflictuel avec les deux actions : $\circlearrowleft T65$ et $\curvearrowright$ T80.

	P. Pren.	Raison	Option	Préoccup.(s)	But(s)
1	Meuniers	"Ne pas prescrire un type de farine unique."	$\circlearrowleft$ T65	Economie	$\nearrow$ Offre segmentée
2	Meuniers	"Les compositions des farines T65 et T80 ne sont pas significativement différentes excepté pour les fibres."	$\circlearrowleft$ T65	Nutrition	$\nearrow$ Oligo-éléments
3	Meuniers	"La teneur en cendres n'est pas un indicateur absolu de la teneur en fibres."	$\curvearrowright$ I.	Technologie	$\nearrow$ Technicité
4	Meuniers	"La production de farine T80 devrait coûter plus cher en raison du mélange des farines."	$\circlearrowleft$ T65	Economie Technologie	$\searrow$ Coûts $\nearrow$ Technicité
5	Meuniers	"83% des consommateurs consomment plus de pain raffiné que de pain complet au regard de leur satiété."	$\circlearrowleft$ T65	Economie	$\nearrow$ Bénéfices
6	Meuniers	"Une hausse de la consommation journalière de pains issus de la farine T65 augmente les apports en fibres."	$\circlearrowleft$ T65	Nutrition	$\nearrow$ Fibres $\nearrow$ Oligo-éléments

TABLE 4.6 – L'argumentation des meuniers, 1^{ere} partie

En réponse à cela, une partie de la raison du dernier argument est réutilisée dans un nouvel argument avancé cette fois-ci par les défenseurs des recommandations (`P.Pren` = PNNS, `But` = $\searrow Sel$, `Action` = $\curvearrowright T80$), voir le tableau 4.7.

Dans certains cas, le principe de précaution peut s'avérer décisif pour le processus décisionnel. Dans cette étude de cas, le risque potentiel de causer un dommage aux consommateurs par le type de pain est négligeable,

	P. Pren.	Raison	Option	Préoccup.(s)	But(s)
4	PNNS	"Une hausse de la consommation journalière de pains issus de la farine T65 augmente les apports en sel dans la diète."	$\curvearrowright$ T80	Nutrition	$\searrow$ Sel

TABLE 4.7 – La réponse argumentée du PNNS

mais les critères sanitaires peuvent parfois être prioritaires sur les critères nutritionnels. Dans le tableau 4.8, sont listés d'autres arguments provenant des meuniers, prenant en compte cette dimension sanitaire (`Preoccup` = Sanitaire), promouvant d'autres buts (`But` = $\searrow$ Mycotoxines, $\searrow$ Résidus de pest., $\searrow$ Acide phytique) et soutenant l'option conservatrice aussi bien que des options plus spécifiques : `Action` = $\circlearrowleft T65$, $\curvearrowright (T80)$ & $D.$ (D. signifie Décortiqué), $\curvearrowright (T80)$ & $B.$ (B. signifie Biologique).

	P. Pren.	Raison	Option	Préoccup.(s)	But(s)
7	Meuniers	"Augmenter le taux d'extraction provoque une hausse des contaminants dans la farine."	$\circlearrowleft$ T65	Sanitaire	$\searrow$ Mycotoxines $\searrow$ Résidus de pest.
8	Meuniers	"Un pré-traitement effectué sur le blé (t.q. le décorticage) peut baisser le niveau de mycotoxines."	$\curvearrowright (T80)$ & $D.$	Sanitaire Technologie	$\searrow$ Mycotoxines $\nearrow$ Technicité
9	Meuniers	"Un pré-traitement accroît les coûts."	$\circlearrowleft$ T65	Economie	$\searrow$ Coûts
10	Meuniers	"Confectionner du pain biologique permet d'éliminer les traces de pesticides."	$\curvearrowright (T80)$ & $B.$	Sanitaire Economie	$\searrow$ Résidus de pest. $\nearrow$ Offre segmentée
11	Meuniers	"Augmenter le taux d'extraction provoque une hausse d'acide phytique"	$\circlearrowleft$ T65	Nutrition	$\searrow$ Acide phytique (A.P)

TABLE 4.8 – L'argumentation des meuniers, 2^{nde} partie

Les boulangers, eux, sont soucieux d'une possible diminution des ventes engendrée par les recommandations de la politique de santé publique. La modification du goût, de la texture et de l'aspect du pain avec la farine T80 pourrait avoir un impact sur l'achat quotidien du pain par les consommateurs. Néanmoins, la sensibilisation sur l'intérêt de ce changement pourrait à contrario avoir un effet positif sur la volonté des consommateurs de payer plus cher un type de baguette française nutritionnellement préférable. Dans le tableau 4.9, sont listés les arguments de la profession boulangère (`P.Pren` = Boulangers), qui prennent en compte certaines préoccupations du consommateur (`Preoccup` = Hédonisme), promouvant d'autres buts (`But` = $\nearrow$ Organoleptique, $\nearrow$ Bénéfices) et soutenant tantôt l'action conservatrice, tantôt l'action réformiste, tantôt des options spécifiques : `Action` = $\circlearrowleft T65$, $\curvearrowright T80$, $\curvearrowright (T80)$ & $L.$ (L. signifie Levain), $\curvearrowright (T80)$ & $T.$ (T. signifie Tradition).

	P. Pren.	Raison	Option	Préoccup.(s)	But(s)
1	Boulangers	"L'utilisation du levain dans la panification permet de dégrader les phytates (en raison d'un faible pH)."	$\curvearrowright (T80)$ & L.	Nutrition	$\searrow$ Acide Phytique
2	Boulangers	"L'acceptabilité pour le pain T80 est un challenge ambitieux (croustillance, goût,...)."	$\circlearrowleft$ T65	Hédonisme Economie	$\nearrow$ Organoleptique $\nearrow$ Offre segmentée
3	Boulangers	"L'acceptabilité pour le pain T80 requiert une adaptation des conditions de panification (fermentation traditionnelle)."	$\curvearrowright (T80)$ & T.	Hédonisme Technologie Economie	$\nearrow$ Organoleptique $\nearrow$ Technicité $\nearrow$ Offre segmentée
4	Boulangers	"L'adaptation du diagramme de panification pour le pain T80 accroît les coûts de fabrication."	$\circlearrowleft$ T65	Economie	$\searrow$ Coûts
5	Boulangers	"Les consommateurs semblent consentir à payer 12% plus cher une baguette labellisée 'source de fibres'."	$\curvearrowright$ T80	Economie	$\nearrow$ Bénéfices

TABLE 4.9 – L'argumentation des boulangers

4.3.3 Le modèle proposé

En premier lieu, il semble important de ne considérer que les arguments concernant une *même* préoccupation pour aboutir à un choix d'option. Ici nous nous concentrons sur la préoccupation nutritionnelle. Par exemple, au cours d'un processus de décision, avant de les évaluer sur une échelle commune, il n'est pas opportun de considérer ensemble un argument nutritionnel et un argument économique. Toutefois plusieurs audiences peuvent légitimement prétendre traiter de la même préoccupation. Pour une préoccupation donnée, nous proposons de considérer une audience comme un *contexte* argumentatif, établissant des préférences entre les arguments exprimés sur cette préoccupation. Une priorisation des buts est définie pour chaque audience.

Définition 4.5 *Un système de décision argumentée étendu* $\mathtt{DF}_{ext}$ *est un tuple* $\langle \mathcal{A}, \mathcal{P}, \mathcal{D}, \mathcal{G}, \trianglerighteq \rangle$ *défini comme suit :*

- $\mathcal{P} = \{p_1, \ldots, p_n\}$ *est un ensemble de préoccupations (nutritionnelle, ...) ;*
- $\mathcal{G} = \mathcal{G}_1, \ldots, \mathcal{G}_n$ *sont des ensembles de buts,* $\mathcal{G}_i$ *étant l'ensemble de buts pertinent pour la préoccupation* p_i *;*
- $\trianglerighteq = \trianglerighteq_1, \ldots, \trianglerighteq_n$ *sont des ensembles de préordres*[1] *complets, chaque* $\trianglerighteq_i$ *défini sur* $\mathcal{G}_i \times \mathcal{G}_i$*, avec* $\trianglerighteq_i = \{\trianglerighteq_i^1, \ldots, \trianglerighteq_i^j\}$*, où* $1 \ldots j$ *sont des contextes et concernent donc chacun une audience particulière (segment de consommateurs, ...) ;*
- $\mathcal{D} = d_1, \ldots, d_m$ *est un ensemble d'actions, équipé d'un ordre partiel* $\angle$ *(appelé relation de spécialisation) et d'une relation d'exclusion mutuelle*

1. Dans un préordre, contrairement à un ordre, certains éléments peuvent être équivalents : on note $(\alpha,\beta) \in \approx_i^k$ si $(\alpha,\beta) \in \trianglerighteq_i^k$ et $(\beta,\alpha) \in \trianglerighteq_i^k$.

notée $\perp$ *;*

• $\mathcal{A} = \mathcal{A}_1, \ldots, \mathcal{A}_n$ *sont des ensembles d'arguments.* $\mathcal{A}_i$ *est l'ensemble d'arguments exprimés dans la préoccupation* p_i*.* $\texttt{but} : \mathcal{A}_i \rightarrow \mathcal{G}_i$ *associe à un argument un but et* $\texttt{action} : \mathcal{A}_i \rightarrow \mathcal{D}$ *associe à un argument une action.*

A partir du système étendu $\texttt{DF}_{ext}$ et pour chaque préoccupation p_i, un système d'argumentation à base de préférences contextuelles $\texttt{CPAF}_i$ peut être extrait, correspondant au débat dans une préoccupation donnée (par exemple nutritionnelle, etc.). Deux arguments y sont en conflit si leurs actions associées sont liées par la relation d'exclusivité mutuelle ou la relation de spécialisation. Les préférences entre arguments s'appuient sur la priorisation des buts associés.

Le système CPAF a été introduit par Amgoud et collab. (2000). Nous définissons donc la correspondance suivante entre un système $\texttt{DF}_{ext}$ et plusieurs modèles CPAF.

Définition 4.6 *A partir d'un* $\texttt{DF}_{ext} = \langle \mathcal{A}, \{p_1, \ldots, p_n\}, \mathcal{D}, \mathcal{G}, \trianglerighteq \rangle$*, un ensemble de systèmes d'argumentation à base de préférences contextuelles* $\{\texttt{CPAF}_1, \ldots, \texttt{CPAF}_n\}$ *peuvent être extraits, où* $\texttt{CPAF}_i = \langle \mathcal{A}_i, \mathcal{R}_i, \succeq_i^1, \ldots, \succeq_i^m \rangle$ *est défini comme suit :*

• $\mathcal{A}_i$ *est l'ensemble des arguments exprimés dans la préoccupation* p_i *;*

• $\mathcal{R}_i \subseteq \mathcal{A}_i \times \mathcal{A}_i$ *est une relation d'attaque telle que* $(\alpha, \beta) \in \mathcal{R}_i$ *si* $\texttt{action}(\alpha) \perp \texttt{action}(\beta)$ *ou si* $\texttt{action}(\alpha) \angle \texttt{action}(\beta)$ *;*

• $\succeq_i^1, \ldots, \succeq_i^m$ *sont des préférences contextuelles (*$\succeq_i^k \subseteq \mathcal{A}_i \times \mathcal{A}_i$*) telles que* $(\alpha, \beta) \in \succeq_i^k$ *si* $\texttt{but}(\alpha) \trianglerighteq_i^k \texttt{but}(\beta)$*.*

Un $\texttt{CPAF}_i$ *peut également être défini comme un tuple* $\texttt{CPAF}_i = \langle \mathcal{A}_i, \texttt{Def}_i^1, \ldots, \texttt{Def}_i^m \rangle$*, où* $\texttt{Def}_i^k \subseteq \mathcal{A}_i \times \mathcal{A}_i$ *est une relation appelée* defeat *telle que* $(\alpha, \beta) \in \texttt{Def}_i^k$ *ssi* $(\alpha, \beta) \in \mathcal{R}_i$ *et* $(\beta, \alpha) \notin \succ_i^k$*.*

Remarque 4.1 *Notons que* $(\alpha, \beta) \in \succ_i^k$ *si* $(\alpha, \beta) \in \succeq_i^k$ *et* $(\beta, \alpha) \notin \succeq_i^k$*.*

Etude de cas suite. Aspects nutritionnels
Le graphe représentant la relation d'attaque non symétrique entre arguments exprimés dans la préoccupation nutritionnelle (voir les tableaux 4.5 à 4.9) est illustré dans la figure 4.3.

Après l'identification des arguments, des buts associés et des actions soutenues, plusieurs audiences peuvent s'exprimer dans chaque préoccupation.

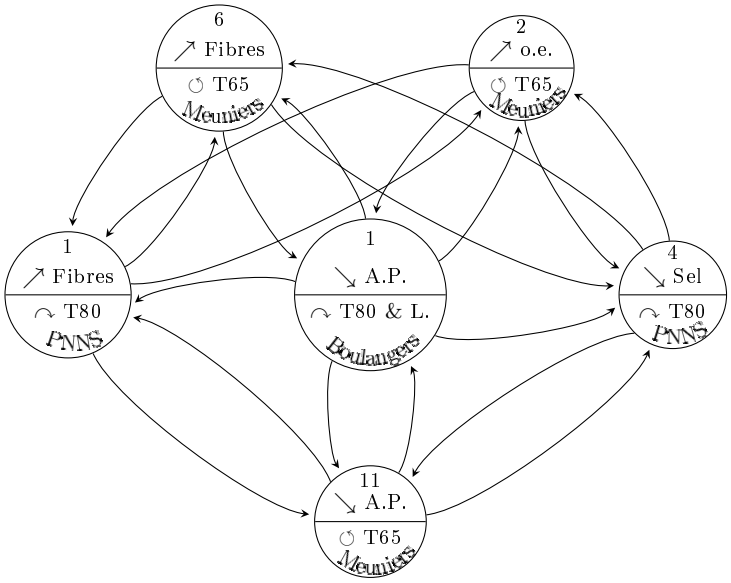

FIGURE 4.3 – Graphe d'attaque pour la préoccupation nutritionnelle

	Description	Audiences
1	Préférence en fibres (i.e. obésité)	$\nearrow$ Fibres $\trianglerighteq^1_N$ $\nearrow$ μnut. $\trianglerighteq^1_N$ $\searrow$ Sel $\approx^1_N$ $\searrow$ A.P.
2	Préférence en micronutriments (i.e. déficit en fer)	$\nearrow$ μnut. $\trianglerighteq^2_N$ $\nearrow$ Fibres $\trianglerighteq^2_N$ $\searrow$ Sel $\approx^2_N$ $\searrow$ A.P.
3	Diminution du sel (i.e. maladies cardiovasculaires)	$\searrow$ Sel $\trianglerighteq^3_N$ $\nearrow$ Fibres $\approx^3_N$ $\nearrow$ μnut. $\approx^3_N$ $\searrow$ A.P.
4	Limitation de l'acide phytique (i.e. végétariens)	$\searrow$ A.P. $\trianglerighteq^4_N$ $\nearrow$ Fibres $\approx^4_N$ $\nearrow$ μnut. $\approx^4_N$ $\searrow$ Sel

TABLE 4.10 – Contextes nutritionnels

Dans ce modèle, les audiences correspondent aux contextes mentionnés plus haut et se caractérisent par une priorisation des buts qui leur est propre. Par exemple, parmi plusieurs buts exprimés par le PNNS, à savoir l'augmentation de la teneur en fibres (du fait de l'"épidémie" d'obésité, de diabète, etc.) et la diminution du taux de sel (pour limiter les atteintes cardiovasculaires, ...), le modèle permet une priorisation différente en fonction de l'audience visée (qui correspond à un contexte). Il permet alors de calculer des actions recommandées pour cette audience.

Etude de cas suite. Audiences en sortie *Nous proposons de caractériser quatre contextes (ou audiences) dans la préoccupation nutritionnelle (no-*

tée p_N), détaillées dans le tableau 4.10. Ces contextes représentent quatre types de consommateurs : les personnes souffrant d'obésité, les personnes présentant une déficience en fer, les personnes atteintes de maladies cardiovasculaires et les personnes ayant un régime végétarien. Ces contextes définissent des priorisations différentes de l'ensemble des buts $\mathcal{G}_N$. La représentation donnée en figure 4.4 décrit le modèle d'argumentation obtenu après l'introduction des contextes du tableau 4.10. Pour chaque contexte, le modèle permet de déterminer si un argument appartient (cercle continu) ou n'appartient pas (cercle en pointillés) à l'ensemble des extensions préférées. Suivant la définition 4.4, le système propose en résultat plusieurs recommandations pour un contexte donné. Notons que le contexte recherchant une diminution du sel fait pencher la balance vers le pain de type T80, tandis que le contexte cherchant à éviter l'acide phytique oriente les résultats vers un pain au levain naturel ou vers le pain actuel de type T65. D'autres audiences sont en faveur du statu quo.

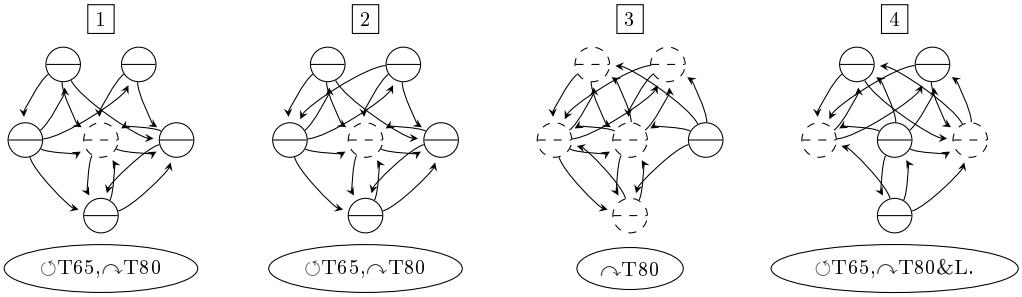

FIGURE 4.4 – Graphes d'attaque propres à chaque audience

4.3.4 Actions recommendées pour d'autres préoccupations et d'autres audiences

Le modèle $\mathtt{DF}_{ext}$ intègre plusieurs préoccupations (sanitaire, technologique, économique et hédoniste). Le tableau 4.11 décrit caractérise plusieurs audiences pour la préoccupation sanitaire et résume les recommandations proposées. Les deux recommandations "pain T80 à base de blé décortiqué"(↷ T80 & D.) et "pain T80 bio" (↷ T80 & B.) peuvent être regroupées en une

seule ($\curvearrowright$ T80 & D.B.), puisque les deux actions sont consistantes l'une vis-à-vis de l'autre. Dans une préoccupation économique, cette recommandation serait contrebalancée par d'autres buts. Par exemple, aucune de ces actions n'est recommandée lorsque la réduction des coûts et l'augmentation des bénéfices sont les buts préférés.

Audience	Action(s) recommandée(s)
$\searrow$ Mycotoxines $\trianglerighteq^1_S$ $\searrow$ Résidus de pesticides	$\circlearrowleft$ T65, $\curvearrowright$ T80 & D.
$\searrow$ Résidus de pesticides $\trianglerighteq^2_S$ $\searrow$ Mycotoxines	$\curvearrowright$ T80 & B.
$\searrow$ Mycotoxines $\approx^3_S$ $\searrow$ Résidus de pesticides	$\circlearrowleft$ T65, $\curvearrowright$ T80 & D., $\curvearrowright$ T80 & B.

TABLE 4.11 – Contextes sanitaires

4.4 Conclusion du chapitre

Comme pour toute action politique, les décideurs s'appuient sur des arguments exprimant différentes préoccupations (santé, économie, service, etc.) afin de recommander une option ayant le plus d'effets positifs pour le plus grand nombre ou pour un public ciblé. Ainsi dans le PNNS, les préoccupations sont essentiellement d'ordre "santé et nutrition", néanmoins il existe d'autres préoccupations telles que les enjeux économiques ou l'acceptabilité hédoniste des consommateurs.

Dans cette étude de cas, la politique de santé publique consiste en une recommandation globale visant à changer le type de farine utilisée dans le pain de consommation courante vendu dans les boulangeries et approvisionnant les restaurations collectives. Les recommandations sont étayées par des arguments exprimant une préoccupation nutritionnelle, principalement liée aux fibres, puis étayée par des arguments économiques liés au rendement de l'extraction des matières premières (ici le blé). Classiquement, l'argumentation vise à transférer à ces recommandations l'adhésion accordée aux raisons. Ainsi, les décideurs, les meuniers, les boulangers et les technologues, qui doivent prendre en compte des préoccupations diverses (sanitaire, hédoniste, etc.), sont engagés dans plusieurs processus d'argumentation, visant à faire évoluer le débat en faveur d'un consensus ($\curvearrowright T80$), d'un compromis ($\curvearrowright$ T80 & Spécification), ou vers un rejet conservateur ($\circlearrowleft T65$).

Comme indiqué dans le présent chapitre, un tel système d'aide à la décision peut aussi être le moyen de cibler un produit alimentaire pour un utilisateur donné (et non pas pour un segment global). Cette méthode "hautement qualitative" ne peut pas être traitée par des approches conventionnelles, en particulier celles de la décision multi-critère bien que des notions telles que les actions (options) ou les objectifs (critères) soient communes aux deux approches. Plusieurs analogies peuvent également être établies dans le domaine médical, où les arguments issus de différentes analyses sont utilisés pour cibler un diagnostic pour les patients. Plus généralement, l'arbitrage fondé sur l'argumentation est une approche prometteuse pour aider les humains à prendre des décisions plus "équilibrées", en tenant compte, par exemple, des trois piliers du concept de durabilité (social, environnemental et économique).

Chapitre 5

Une méthode d'ingénierie inverse pour le pilotage de filière

L'évaluation de la qualité alimentaire est un processus complexe car il repose sur de nombreux critères historiquement regroupés en quatre grands types : qualité nutritionnelle, sensorielle, de praticité et d'hygiène. Ceux-ci peuvent être complétés par d'autres préoccupations émergentes telles l'impact environnemental, les phénomènes économiques, etc. Toutefois, tous ces aspects de la qualité et leurs différentes composantes ne sont pas toujours compatibles et leur amélioration simultanée est un problème qui n'admet pas de solution évidente, ce qui constitue un véritable verrou pour la prise de décision. Ce chapitre propose une méthode d'aide à la décision en ingénierie inverse, c'est-à-dire guidée par les objectifs à l'aval de la filière. Elle se traduit par une approche en chaînage arrière s'appuyant sur l'argumentation.

Ce travail a été réalisé en collaboration avec Madalina Croitoru et publié dans (Thomopoulos et Croitoru, 2013). Son implémentation est en cours dans le cadre du stage d'Ahmed Chadli.

5.1 Problématique

L'offre au sein d'une filière agroalimentaire se construit traditionnellement des producteurs vers les consommateurs, en passant par les étapes intermédiaires de transformation, stockage, transport, conditionnement ..., c'est-à-dire de l'amont vers l'aval. Plus récemment, les contraintes de qualité

s'accroissant, la notion de pilotage de filière par l'aval a émergé. C'est cette fois la demande, et non l'offre, qui fixe le cahier des charges des produits et c'est à la filière de s'adapter et de trouver les moyens d'y répondre. Les méthodes permettant, en partant de critères-cibles souhaités pour le produit final, d'identifier des moyens d'y parvenir en termes d'intervention sur les différentes étapes de la filière, sont qualifiées d' "ingénierie inverse".

L'ingénierie inverse est connue au sein des filières pour poser de réels verrous méthodologiques à sa mise en place. Pour quelles raisons ? Du point de vue des systèmes d'information, il n'existe pas réellement d'obstacle à retrouver l'information concernant les moyens de parvenir à une fin, si celle-ci est connue dans l'autre sens, le sens chronologique de la filière, des procédés vers le résultat. Les verrous sont en fait de deux types : le premier est la difficulté à définir le cahier des charges escompté pour le produit fini, les critères de qualité étant multiples, discutables et pas nécessairement compatibles entre eux. Comment dans ces conditions fixer des priorités pour les caractéristiques du produit fini ? La seconde difficulté réside dans le fait que l'impact des différentes étapes d'une filière sur le produit final n'est pas complètement connu, même dans le sens de déroulement de la chaîne de production. Certaines étapes sont plus étudiées que d'autres, plusieurs étapes successives peuvent avoir des effets opposés dont on maîtrise mal l'effet résultant, les critères-cibles peuvent être en-dehors des caractéristiques des produits issus des filières traditionnelles. Tous ces éléments constituent des freins à la mise en place de l'ingénierie inverse.

Dans ce chapitre, nous abordons deux questions de cette problématique. Tout d'abord nous acceptons l'idée selon laquelle un cahier des charges unique ne peut pas être établi et que plusieurs points de vue complémentaires, éventuellement contradictoires, peuvent être exprimés (nutritionnel, environnemental, gustatif, etc.). Il s'agit ensuite d'évaluer leur compatibilité (ou incompatibilité) et d'identifier des solutions satisfaisant un ensemble maximal de points de vue.

Pour atteindre cet objectif, une bonne connaissance de l'organisation des filières agroalimentaires est nécessaire. Celles-ci ont évolué en quelques décennies en une industrie dynamique tournée vers l'innovation, afin de répondre à la demande en produits à la fois sûrs, sains et attractifs. Elles s'appuient sur les connaissances issues de leur savoir-faire mais aussi sur les résultats de la recherche, qui s'est construite essentiellement sur une approche analytique, fondée sur l'expérimentation et généralisée par des modèles mathé-

matiques. Ces modèles, lorsqu'ils sont disponibles, peuvent alors être utilisés en simulation. Toutefois ils ne traitent pas la question de l'ingénierie inverse qui permet de partir de la demande. De plus l'arbitrage à l'échelle de la filière, contrairement à cette approche, est de nature qualitative, sous peine de s'avérer trop partielle et spécifique. Nous faisons donc l'hypothèse ici qu'une démarche qualitative en ingénierie inverse est possible. Nous proposons un cadre logique s'appuyant sur l'argumentation et introduisons une méthode de décision mettant en œuvre le chaînage arrière. L'approche, totalement générique, est illustrée dans le cas de la filière boulangère, suite au débat autour du changement de la teneur en cendres de la farine utilisée pour le pain de consommation courante, ainsi que de la teneur en sel, dans le cadre des recommandations du PNNS (Programme National Nutrition-Santé) émanant du Ministère de la Santé.

Le chapitre est organisé de la façon suivante. La partie 5.2 introduit les éléments du formalisme logique. La partie 5.3 présente la modélisation du problème. La partie 5.4 expose la démarche d'aide à la décision. Les parties 5.5 et 5.6 proposent une discussion et concluent.

5.2 Les éléments du formalisme

5.2.1 Pourquoi un langage logique ?

Nous attendons du langage de représentation utilisé qu'il permette d'effectuer des raisonnements : à partir des connaissances initiales, explicitement représentées, produire des connaissances et en particulier ici faire de la déduction, répondre à une requête, calculer des contradictions. Le choix d'un langage logique permet à de tels raisonnements d'être bien-fondés.

Par rapport à la logique des propositions, la logique du premier ordre permet de structurer les énoncés utilisés en exprimant des relations entre entités : elle fait apparaître les notions de prédicat, de constante, de variable et introduit des quantificateurs ("pour tout", "il existe"). Cette structuration permet d'affiner la représentation des connaissances et les raisonnements de façon utile à l'application : on peut ainsi représenter un pain comme une entité, exprimer des propriétés valables pour tous les pains, etc. Des exemples seront donnés tout au long des parties suivantes.

Dans la suite de cette partie, nous introduisons les principales notions de

notre langage logique suivant les notations de (Chein et Mugnier, 2009). Ce langage correspond à la partie logique du formalisme présenté dans le chapitre 2. Nous rappelons ici les notions utilisées pour la cohérence du présent chapitre. Notons que nous n'utilisons pas de règles introduisant des variables existentielles en conclusion, pour deux raisons. D'une part, parce que ce type de règle ne s'est jamais présenté dans le cas d'étude considéré. D'autre part, parce que de telles règles compliqueraient les dérivations inverses utilisées dans ce chapitre et présentées dans la partie suivante.

5.2.2 Définitions en logique du premier ordre

Soit $\mathbf{C}$ un ensemble de constantes et $\mathbf{P} = P_1 \cup P_2 \ldots \cup P_n$, où les P_i sont des ensembles de prédicats d'arité commune i. Soit $\mathbf{V}$ un ensemble infini dénombrable de *variables*. Nous définissons l'ensemble des *termes* par $\mathbf{T} = \mathbf{V} \cup \mathbf{C}$. Suivant la dénomination usuelle, étant donnés $i \in \{1 \ldots n\}$, $p \in P_i$ et $t_1, \ldots, t_i \in \mathbf{T}$ nous appelons $p(t_1, \ldots, t_i)$ un *atome*. Soit γ une conjonction d'atomes. Nous notons $var(\gamma)$ l'ensemble des variables de γ. Etant donnés $\mathbf{V}, \mathbf{C}$ et $\mathbf{P}$, un *fait* sur $\mathbf{V}$ est la fermeture existentielle d'une conjonction d'atomes sur $\mathbf{V}$ (une $\exists\wedge$formule). Etant donné un ensemble de faits S, nous notons $\bigwedge S$ la fermeture existentielle de la conjonction de tous les faits de S. Soulignons qu'il n'y a ni négation ni disjonction dans les faits.

Une *interprétation* de $(\mathbf{P}, \mathbf{C})$ est un couple $I = (\triangle, .^I)$ où $\triangle$ est le domaine d'interprétation (éventuellement infini) et $.^I$, la fonction d'interprétation, satisfait (a) $\forall c \in \mathbf{C}, c^I \in \triangle$, (b) $\forall i, \forall p \in P_i, p^I \subseteq \triangle^i$ et (c) $\forall (c, c')$ constantes distinctes de $\mathbf{C}$, $c^I \neq c'^I$.

Exemple 5.1 *Pain, Céréale, PeuSalé, SansContaminants sont des exemples de prédicats unaires (d'arité 1) et EstIngrédient un exemple de prédicat binaire (d'arité 2) dans l'application.*
blé, avoine, seigle, orge sont des exemples de constantes.
Céréale(blé) est un atome.
$\exists$ *x (Pain(x)* $\wedge$ *EstIngrédient(blé, x)) est un exemple de fait.*

5.2.3 Conséquence logique, substitution et homomorphisme

Nous notons $\models$ la *conséquence* logique classique de la logique du premier ordre.

Etant donnés un ensemble de variables $\mathbf{X}$ et un ensemble de termes $\mathbf{T}$, une substitution σ de $\mathbf{X}$ par $\mathbf{T}$ est une application de $\mathbf{X}$ dans $\mathbf{T}$ (notée $\sigma : \mathbf{X} \rightarrow \mathbf{T}$). Etant donné une conjonction d'atomes γ, $\sigma(\gamma)$ désigne la conjonction d'atomes obtenue à partir de γ en remplaçant chaque occurrence de $x \in \mathbf{X} \cap var(\gamma)$ par $\sigma(x)$. Si un fait F est la fermeture existentielle d'une conjonction d'atomes γ alors $\sigma(F)$ est la fermeture existentielle de $\sigma(\gamma)$. Enfin, un homomorphisme d'un fait F vers un fait F' est une substitution σ de $var(F)$ par un sous-ensemble des termes de F' tel que $\sigma(F) \subseteq F'$. L'homomorphisme peut aussi être défini sur les hypergraphes correspondant aux faits. Le théorème fondamental (Chein et Mugnier, 2009) établit que $F' \models F$ si et seulement si il existe un homomorphisme de F dans F'.

Exemple 5.2 *(Pain(banette) ∧ EstIngrédient(blé, banette)) est une substitution de (Pain(x) ∧ EstIngrédient(blé, x)), où banette est une constante.*

5.2.4 Règles et dérivation

Définissons maintenant les règles qui permettront d'enrichir les faits ci-dessus par de nouveaux faits (si elles sont applicables). Une règle R est une formule $\forall X \forall Y \ (H[X,Y] \rightarrow C[Y])$ où H, l'hypothèse, et C, la conclusion, sont deux conjonctions d'atomes ; X et Y sont les ensembles de variables apparaissant respectivement dans H et dans C.

Exemple 5.3 *Un exemple de règle est le suivant :*
∀ x (Pain(x) ∧ SansPesticides(x) ∧ SansMycotoxines(x) → SansContaminants(x)).

Une règle $R = (H, C)$ est applicable à un fait F s'il existe un homomorphisme π de H dans F. Dans ce cas, l'application de R dans F conformément à π produit un fait $F \cup \pi(C)$. On dit alors que le nouveau fait (qui inclut l'ancien et y ajoute de l'information nouvelle) est une dérivation immédiate de F par R, également notée abusivement $R(F)$.

Notons que cette technique est couramment utilisée, par exemple, pour la réponse à une requête en chaînage arrière (Van Melle et collab., 1981; Clocksin et Mellish, 1984; Baget et Salvat, 2006), où une requête est réécrite conformément aux règles (puis une occurrence de la requête réécrite est recherchée dans les faits de la base de connaissances). Le même mécanisme est également abordé par les algorithmes de raisonnement abductif (Klarman et collab., 2011) où un ensemble de faits minimal (au sens de l'inclusion) doit être ajouté à la base de connaissances pour pouvoir déduire la requête.

Dans ce chapitre nous ne nous intéressons pas nécessairement à la minimalité de ces ensembles, du fait de la petite taille des bases de connaissances employées. De plus, la sémantique obtenue par les extensions du système d'argumentation est similaire. Ce point pourrait toutefois constituer un fil conducteur intéressant pour les travaux à venir, afin d'optimiser notre méthode à des cas d'utilisation plus larges.

Exemple 5.4 *Soit :*
$F =$ *Pain(bleuette)* $\wedge$ *SansPesticides(bleuette)* $\wedge$ *SansMycotoxines(bleuette)*
et R *la règle de l'Exemple 5.3.*

R est applicable à F et produit par dérivation le nouveau fait suivant :
Pain(bleuette) $\wedge$ *SansPesticides(bleuette)* $\wedge$ *SansMycotoxines(bleuette)* $\wedge$ *SansContaminants(bleuette).*

De façon similaire, une règle $R = (H, C)$ est inversement applicable à un fait F s'il existe un homomorphisme π de C dans F. Dans ce cas, l'application inverse de R dans F conformément à π produit un nouveau fait F' tel que $R(F') = F$. On dit alors que le nouveau fait est dérivation inverse immédiate de F par R, également notée abusivement $R^{-1}(F)$.

Exemple 5.5 *Soit :*
$F =$ *Pain(bleuette)* $\wedge$ *SansContaminants(bleuette)*
et R *la règle de l'Exemple 5.3.*

R est inversement applicable à F et produit par dérivation inverse :
$F' =$ *Pain(bleuette)* $\wedge$ *SansPesticides(bleuette)* $\wedge$ *SansMycotoxines(bleuette).*

Soit F un fait et $\mathcal{R}$ un ensemble de règles. Un fait F' est appelé $\mathcal{R}$-dérivation de F s'il existe une séquence finie (appelée séquence de dérivation) $F = F_0, F_1, \ldots, F_k = F'$ telle que pour tout $1 \leq i \leq k$ il existe une règle

$R = (H, C) \in \mathcal{R}$ telle que F_i est une dérivation immédiate de F_{i-1} par R. L'ensemble de tous les faits obtenus à partir d'un fait F et d'un ensemble de règles $\mathcal{R}$ par toutes les $\mathcal{R}$-dérivations est noté $\mathcal{R}(F)$. De même, partant d'un ensemble de faits $\mathcal{F}$, l'ensemble de tous les faits obtenus à partir de $\mathcal{F}$ et de $\mathcal{R}$ par toutes les $\mathcal{R}$-dérivations est noté $\mathcal{R}(\mathcal{F})$. Soulignons que, étant donnée la forme des règles, les dérivations ne sont pas infinies.

De façon similaire, nous pouvons définir des $\mathcal{R}$-dérivations inverses de F et, étant donné F et un ensemble de règles $\mathcal{R}$, toutes les $\mathcal{R}$-dérivations inverses de F (noté $\mathcal{R}^{-1}(F)$).

Etant donnée la forme des règles dans le cas d'étude, ainsi que le nombre limité de faits et de règles, nous n'optimisons pas ce processus par une étude de minimalité (comme (Klarman et collab., 2011)). Nous projetons d'approfondir ces aspects par la suite comme discuté dans la dernière partie du chapitre.

5.2.5 Expression de l'inconsistance

Nous considérons une forme particulière de négation exprimée au moyen de contraintes négatives : elle permet d'exprimer une information qui ne doit pas pouvoir se déduire de la base de connaissances. L'objectif des contraintes négatives est d'exprimer des inconsistances dans les faits (y compris des faits partiellement générés par des règles). Plus précisément, une contrainte est une formule $\neg N$ où N est un fait ($\exists\wedge$formule). Un ensemble de contraintes négatives noté $\mathcal{N}$ est un ensemble de négations de faits ($\exists\wedge$formules). Notons qu'une contrainte négative peut aussi être vue comme une règle de la forme $(N, \bot)$. Une contrainte négative est satisfaite par un fait F si $F \not\models N$. Si un fait satisfait une contrainte négative alors il est *consistant*. Sinon le fait est *inconsistant*.

Exemple 5.6 *$N = \exists\ x$ (Augmentation(x) $\wedge$ Diminution(x)) est un exemple de contrainte négative.*

Une contrainte négative est satisfaite par un fait F et un ensemble de règles $\mathcal{R}$ si $\mathcal{R}(F) \not\models N$. Dans ce cas on dit que le fait est $\mathcal{R}$-*consistant* (sinon il est $\mathcal{R}$-*inconsistant*). De même un ensemble de faits $\mathcal{F}$ est dit $\mathcal{R}$-*consistant* si $\mathcal{R}(\mathcal{F}) \not\models N$ (sinon il est $\mathcal{R}$-*inconsistant*).

5.2.6 Base de connaissances consistante

Une base de connaissances sur un vocabulaire est un triplet $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$ composé de trois ensembles finis de formules :

- un ensemble $\mathcal{F}$ de *faits*, représentant des connaissances assertionnelles positives ;
- un ensemble $\mathcal{R}$ de *règles*, représentant des connaissances ontologiques pouvant être appliquées aux faits pour obtenir de nouveaux faits ;
- un ensemble $\mathcal{N}$ de négations de $\exists\wedge$formules, représentant des contraintes négatives que les faits doivent satisfaire avant ou après l'application des règles.

Une telle base de connaissances est *consistante* si $(\mathcal{F}, \mathcal{R})$ satisfait chaque contrainte de $\mathcal{N}$ (tous les faits de $\mathcal{F}$ sont $\mathcal{R}$-consistants).

Exemple 5.7 *Soit $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$ avec :*

- *$\mathcal{F}$ constitué des faits suivants :*

$F_1 =$ *Pain(bleuette)* $\wedge$ *SansContaminants(bleuette)*
$F_2 = \exists$ *e TauxExtraction(e,bleuette)*
$F_3 = \exists$ *f (TeneurFibres(f,bleuette)* $\wedge$ *Elevé(f))*

- *$\mathcal{R}$ constitué des trois règles suivantes :*

$R_1 = \forall$ *x,y (Pain(x)* $\wedge$ *TauxExtraction(y,x)* $\wedge$ *SansPesticides(x)* $\rightarrow$ *Diminution(y))*
$R_2 = \forall$ *x,y,z (Pain(x)* $\wedge$ *TauxExtraction(y,x)* $\wedge$ *TeneurFibres(z,x)* $\wedge$ *Elevé(z)* $\rightarrow$ *Augmentation(y))*
$R_3 = \forall$ *x (Pain(x)* $\wedge$ *SansContaminants(x)* $\rightarrow$ *SansPesticides(x)* $\wedge$ *SansMycotoxines(x))*

- *$\mathcal{N}$ contenant la contrainte négative suivante :*

$N = \exists$ *x (Augmentation(x)* $\wedge$ *Diminution(x))*

$\mathcal{K}$ est inconsistant car $(\mathcal{F}, \mathcal{R}) \models N$. En effet F_1 et R_3 permettent de déduire SansPesticides(bleuette). Combiné à F_2 et R_1 on obtient Diminution(e). Or F_3 et R_2 conduisent à Augmentation(e), ce qui viole la contrainte négative N.

5.2.7 Réponse à une requête : chaînage avant et chaînage arrière

Un fait Q est *dérivé* de $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$ si et seulement si, soit $\mathcal{K}$ est inconsistant (et tout peut alors en être déduit), soit $(F, \mathcal{R}) \models Q$. Si Q représente une requête, ce dernier cas signifie qu'il existe une réponse à la requête Q dans $\mathcal{K}$.

Répondre à une requête Q (représentée comme un fait dans le formalisme) relève de deux approches algorithmiques différentes : le chaînage avant ou le chaînage arrière. Les deux approches consistent respectivement à :

1. trouver une réponse à Q dans les $\mathcal{R}$-dérivations des faits de la base de connaissances ;
2. calculer les $\mathcal{R}$-dérivations inverses de la requête et déterminer s'il existe une substitution dans les faits.

C'est cette seconde approche qui va nous intéresser par la suite.

5.3 Modéliser le problème

5.3.1 Présentation du cas d'étude

Le cas d'étude est celui exposé en détail dans le chapitre 4. Nous en rappelons ici les principales lignes. Ce cas d'étude concerne initialement le débat autour du changement de la teneur en cendres de la farine utilisée pour le pain de consommation courante (pain français). Différents acteurs de la filière sont concernés, notamment le Ministère de la Santé au travers de ses recommandations dans le cadre du PNNS (Programme National Nutrition-Santé), les meuniers, les boulangers, les nutritionnistes et les consommateurs.

En effet, le PNNS recommande de privilégier les produits céréaliers complets et en particulier de passer à un pain de consommation courante de type T80, i.e. fabriqué à partir de farine contenant un taux de cendres (taux de matière minérale) de 0.8 %, au lieu du type T65 (0.65 % de matière minérale) actuellement utilisé. Augmenter le taux de cendres revient effectivement à utiliser une farine plus complète, la matière minérale étant concentrée dans les couches périphériques du grain de blé, de même que bon nombre de constituants d'intérêt nutritionnel (vitamines, fibres, ...). Toutefois, les couches

périphériques du grain sont aussi les plus exposées aux produits phytosanitaires, ce qui ne les rend pas recommandables d'un point de vue sanitaire, à moins d'utiliser de la farine biologique.

D'autres arguments, de différentes natures, sont en faveur ou défaveur d'un pain plus complet. D'un point de vue organoleptique par exemple, celui-ci perd en "croustillant". D'un point de vue nutritionnel, l'argument selon lequel les fibres sont bénéfiques pour la santé est discuté, certaines d'entre elles étant irritantes pour l'appareil digestif. D'un point de vue économique, les boulangers craignent une perte de bénéfices en vendant moins, car un pain plus complet augmente la satiété – ce qui est bénéfique d'un point de vue nutritionnel, pour la régulation de l'appétit et pour la lutte contre les déséquilibres et pathologies alimentaires. Cependant un pain plus complet nécessite également moins de farine et plus d'eau pour sa confection, ce qui en diminue le coût. Les meuniers craignent pour leur part une perte de technicité dans la confection des farines.

Au-delà de la polémique sur le choix entre deux alternatives (T65 ou T80), il s'agit ici de considérer le problème dans son ensemble. On peut ainsi distinguer les différents points de vue en jeu, identifier les caractéristiques-cibles souhaitables, estimer les moyens d'y parvenir, sans se restreindre à la problématique du type de farine, et déterminer les différentes options qui se dessinent.

5.3.2 Exprimer les caractéristiques-cibles suivant différents points de vue

Exprimer les caractéristiques-cibles (ou buts) suivant différents points de vue consiste, d'une part, à identifier les différentes facettes intervenant dans la construction de la qualité du produit : les points de vue, grandes catégories de critères en jeu telles que nutrition, qualité hédonique, environnement, technologie, etc. ; d'autre part, à les décliner selon leurs différentes composantes (critères tels que le taux ou la qualité des fibres, minéraux, vitamines, etc.) et définir les orientations souhaitables, c'est-à-dire les valeurs souhaitées pour ces critères. Ces valeurs sont exprimées de façon qualitative, indépendamment les unes des autres dans un premier temps.

Ces premières étapes ont été réalisées, dans notre étude, à partir des sources d'information détaillées dans le chapitre précédent.

FIGURE 5.1 – Buts nutritionnels

Un extrait des résultats obtenus dans la filière boulangère est synthétisé par les figures 5.1, pour les aspects nutritionnels, et 5.2, pour les aspects organoleptiques. Tous les points de vue ne sont pas abordés ici par souci de simplicité.

5.3.3 Formalisation des buts

Soit $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$ une base de connaissances consistante. Il s'agit en l'occurrence de la base de connaissances partagée et admise par l'ensemble des acteurs impliqués dans la construction de la qualité des produits de la filière (meuniers, boulangers, consommateurs, nutritionnistes, technologues, chercheurs, pouvoirs publics). Nous supposons ici que les règles et les contraintes négatives sont partagées. Cette hypothèse sera relâchée lors de travaux futurs.

Les buts des différents acteurs peuvent être vus comme un ensemble de conjonctions d'atomes fermées existentiellement. Nous les notons $G_1, G_2, \ldots, G_n$.

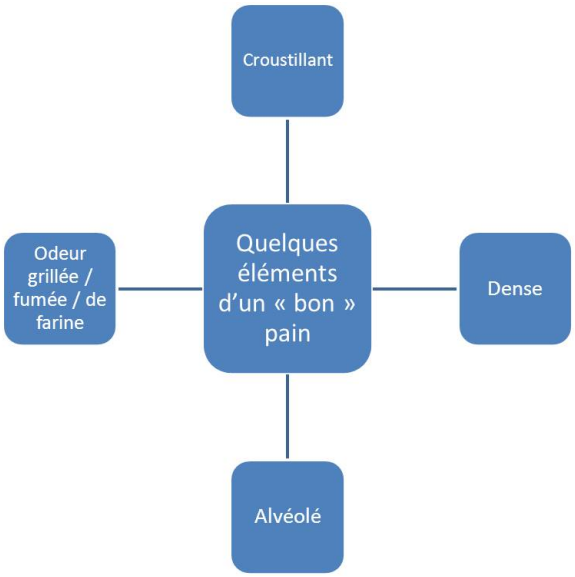

FIGURE 5.2 – Buts organoleptiques

Soit $\mathcal{G}$ l'ensemble des buts $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$. Les buts correspondent, par une fonction $\kappa : \mathcal{G} \rightarrow 2^{\mathcal{V}}$, à un ensemble de points de vue $\mathcal{V}$ (nutritionnel, organoleptique, etc.). Cette fonction peut assigner un but à un ou plusieurs points de vue et chaque point de vue peut être associé à un ou plusieurs buts. Etant donné un but G_i, l'ensemble des points de vue associés à ce but est noté $\kappa(G_i)$. De façon similaire, étant donné un point de vue v_i, l'ensemble des buts associés est noté $\kappa^{-1}(v_i)$.

Exemple 5.8 *Soit l'ensemble de points de vue* $\mathcal{V} = \{$*nutritionnel, sanitaire, organoleptique*$\}$ *et* $\mathcal{G}$ *constitué des buts suivants :*
$G_1 = \exists\ x\ (Pain(x) \wedge PeuSalé(x))$
$G_2 = \exists\ x\ (Pain(x) \wedge SansContaminants(x))$
$G_3 = \exists\ x\ (Pain(x) \wedge Croustillant(x))$
$G_4 = \exists\ x\ (Pain(x) \wedge RicheOligoéléments(x))$

On a $\kappa(G_1) = \kappa(G_4) = nutritionnel$, $\kappa(G_2) = sanitaire$ *et* $\kappa(G_3) = organoleptique$.

Inversement $\kappa^{-1}(nutritionnel) = \{G_1, G_4\}$, $\kappa^{-1}(sanitaire) = \{G_2\}$ *et* $\kappa^{-1}(organoleptique) = \{G_3\}$.

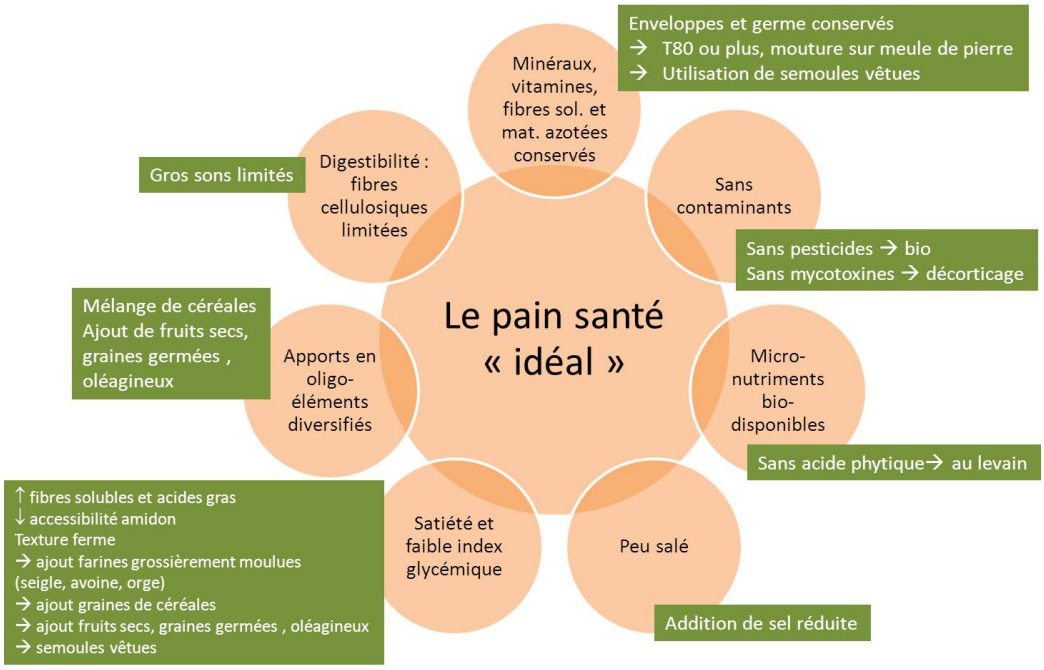

FIGURE 5.3 – Moyens d'atteindre les buts nutritionnels

5.3.4 Traduire l'ingénierie inverse

L'ensemble de règles $\mathcal{R}$ de la base de connaissances $\mathcal{K}$ contient un ensemble de conditions suffisantes pour atteindre les buts G_i.

L'illustration sur la filière boulangère est schématisée par la figure 5.3.4 pour les aspects nutritionnels.

Exemple 5.9 *Pour atteindre le but* $G_1 = \exists\ x\ (Pain(x) \wedge PeuSalé(x))$, *la base de connaissances* $\mathcal{K}$ *contient la règle suivante, qui exprime une condition suffisante :*
$\forall\ x,y\ (Pain(x) \wedge AdditionSel(y,x) \wedge Diminution(y) \rightarrow PeuSalé(x))$

5.3.5 Formalisation du processus d'ingénierie inverse

Soit G_i un but et $\mathcal{K}$ la base de connaissances. $\mathcal{K}$ est consistant et G_i ne peut pas être dérivé de $\mathcal{K}$. Nous calculons les $\mathcal{R}$-dérivations inverses de G_i (où $\mathcal{R}$ est l'ensemble des règles de la base de connaissances). Nous ajoutons tous

les $\mathcal{R}^{-1}(G_i)$ aux faits. Nous obtenons ainsi une nouvelle base de connaissances $\mathcal{K}_i$ qui diffère de $\mathcal{K}$ uniquement par son ensemble de faits (qui maintenant inclut aussi $\mathcal{R}^{-1}(G_i)$) : $\mathcal{K} = (\mathcal{F} \cup \mathcal{R}^{-1}(G_i), \mathcal{R}, \mathcal{N})$. Nous imposons également que $\mathcal{K}_i$ soit consistant.

Dans la suite du chapitre nous simplifierons la notation de $\mathcal{F} \cup \mathcal{R}^{-1}(G_i)$ par $\mathcal{F}_i$ (l'ensemble des faits obtenus concernant le but G_i).

Considérons l'ensemble des buts $\mathcal{G} = \{G_1, G_2, \ldots, G_n\}$ et la base de connaissances initiale $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$. Comme décrit ci-dessus nous calculons les n bases de connaissances correspondant à chaque but : $\mathcal{K}_i = (\mathcal{F} \cup \mathcal{R}^{-1}(G_i), \mathcal{R}, \mathcal{N})$ pour chaque $i = 1, \ldots, n$. Considérons maintenant l'union de toutes ces bases de connaissances :

$$\mathcal{K}_{agg} = (\mathcal{F} \bigcup_{i=1,\ldots,n} \mathcal{R}^{-1}(G_i), \mathcal{R}, \mathcal{N})$$

5.3.6 Exemple illustratif

Illustrons la démarche sur l'exemple suivant.

Exemple 5.10 *Soit $\mathcal{K} = (\mathcal{F}, \mathcal{R}, \mathcal{N})$ avec :*

- $\mathcal{F} = \{F_1\} = \{$*TauxExtractionCourant(T65)*$\}$
- $\mathcal{R}$ *constitué des règles suivantes :*

$R_1 =$ $\forall$ *x,y (Pain(x)* $\wedge$ *TauxExtraction(y,x)* $\wedge$ *Diminution(y)* $\rightarrow$ *Digeste(x))*
$R_2 =$ $\forall$ *x,z (Pain(x)* $\wedge$ *AdditionSel(z,x)* $\wedge$ *Diminution(z)* $\rightarrow$ *PeuSalé(x))*
$R_3 =$ $\forall$ *x,y (Pain(x)* $\wedge$ *TauxExtraction(y,x)* $\wedge$ *Augmentation(y)* $\rightarrow$ *RicheOligoéléments(x))*
$R_4 =$ $\forall$ *x,y (Pain(x)* $\wedge$ *TauxExtraction(y,x)* $\wedge$ *Diminution(y)* $\rightarrow$ *SansPesticides(x))*

- $\mathcal{N}$ *contenant la contrainte négative suivante :*

$N =$ $\exists$ *x (Augmentation(x)* $\wedge$ *Diminution(x))*

Soit $\mathcal{G}$ constitué des buts suivants :

$G_1 =$ $\exists$ *p (Pain(p)* $\wedge$ *Digeste(p))* *avec* $\kappa(G_1) =$ *nutritionnel*
$G_2 =$ $\exists$ *p (Pain(p)* $\wedge$ *PeuSalé(p))* *avec* $\kappa(G_2) =$ *nutritionnel*
$G_3 =$ $\exists$ *p (Pain(p)* $\wedge$ *RicheOligoéléments(p))* *avec* $\kappa(G_3) =$ *nutritionnel*
$G_4 =$ $\exists$ *p (Pain(p)* $\wedge$ *SansPesticides(p))* *avec* $\kappa(G_4) =$ *sanitaire.*

On a $\mathcal{K}_1 = (\mathcal{F}_1, \mathcal{R}, \mathcal{N})$ avec $\mathcal{F}_1 = \mathcal{F} \cup \mathcal{R}^{-1}(G_1)$ constitué des deux faits suivants :

F_1 = *TauxExtractionCourant(T65)*
F_2 = *Pain(p)* $\wedge$ *TauxExtraction(τ,p)* $\wedge$ *Diminution(τ)*

$\mathcal{K}_2 = (\mathcal{F}_2, \mathcal{R}, \mathcal{N})$ *avec* $\mathcal{F}_2 = \mathcal{F} \cup \mathcal{R}^{-1}(G_2)$ *constitué des deux faits suivants :*
F_1 = *TauxExtractionCourant(T65)*
F_3 = *Pain(p)* $\wedge$ *AdditionSel(s,p)* $\wedge$ *Diminution(s)*

$\mathcal{K}_3 = (\mathcal{F}_3, \mathcal{R}, \mathcal{N})$ *avec* $\mathcal{F}_3 = \mathcal{F} \cup \mathcal{R}^{-1}(G_3)$ *constitué des deux faits suivants :*
F_1 = *TauxExtractionCourant(T65)*
F_4 = *Pain(p)* $\wedge$ *TauxExtraction(τ,p)* $\wedge$ *Augmentation(τ)*

$\mathcal{K}_4 = (\mathcal{F}_4, \mathcal{R}, \mathcal{N})$ *avec* $\mathcal{F}_4 = \mathcal{F} \cup \mathcal{R}^{-1}(G_4)$ *constitué des deux faits suivants :*
F_1 = *TauxExtractionCourant(T65)*
F_2 = *Pain(p)* $\wedge$ *TauxExtraction(τ,p)* $\wedge$ *Diminution(τ)*

On obtient $\mathcal{K}_{agg} = (\mathcal{F} \bigcup_{i=1,\ldots,n} \mathcal{R}^{-1}(G_i), \mathcal{R}, \mathcal{N})$ *avec* $\mathcal{F} \bigcup_{i=1,\ldots,n} \mathcal{R}^{-1}(G_i) = \{F_1, F_2, F_3, F_4\}$.

Comme le montre l'exemple, il peut arriver que $\mathcal{K}_{agg}$ soit inconsistant (ainsi F_2, F_4 violent la contrainte négative N). C'est le cas même pour des buts appartenant au même point de vue (ainsi F_2 est obtenu pour G_1 et G_4, associés respectivement aux points de vue nutritionnel et sanitaire, et F_4 est obtenu pour G_3, associé au point de vue nutritionnel).

Afin d'isoler des sous-ensembles maximaux consistants nous utilisons l'argumentation (Dung, 1995) qui, au moyen des extensions, permet de les calculer. De plus, les extensions nous permettront de voir quels points de vue sont associés à chaque sous-ensemble maximal cohérent (au moyen de la fonction κ). On pourra alors utiliser soit des procédures de vote simple pour déterminer le point de vue à suivre, soit d'autres modes de sélection à base de préférences. Les notions utiles sont rappelées dans la section 5.4 car différentes de celles du chapitre 4.

5.4 Aide à la décision

Nous venons de voir qu'un ensemble de buts a été explicité. Puis un ensemble de faits, identifiés comme des conditions suffisantes pour atteindre ces buts, a été calculé. Il nous faut à présent définir les éléments d'un système

d'argumentation (Dung, 1995) dans ce contexte, et en particulier ce qu'est un argument et ce qu'est une attaque. Ils serviront de base dans le processus d'aide à la décision.

5.4.1 Calcul des arguments et des extensions

Nous partons de $\mathcal{K}_{agg}$, l'union $\mathcal{R}$-inconsistante des n bases de connaissances respectant chacune le même ensemble de règles et de contraintes mais ayant des faits $\mathcal{R}$-consistants différents.

Un *argument* a sur $\mathcal{K}_{agg}$ est une séquence $a = (\varphi_0, \varphi_1, \ldots, \varphi_k)$ où φ_0 est un sous-ensemble de faits de $\mathcal{K}_{agg}$, $\varphi_1 = \bigwedge \varphi_0$, tous les φ_i $(i = 1, \ldots, k)$ sont des faits $\mathcal{R}$-consistants et, pour tout $i = 2, \ldots, k$, $\exists R \in \mathcal{R}$ une règle telle que $\varphi_{i+1} = R(\varphi_i)$, i.e. φ_{i+1} est une dérivation immédiate de φ_i.

Exemple 5.11 *Reprenons la base de connaissances $\mathcal{K}_{agg}$ de l'exemple 5.10. Un exemple d'argument est :*
$a = (\{F_2\}, F_2, R_1(F_2))$
avec $R_1(F_2) =$ *Pain(p)* $\wedge$ *TauxExtraction(*τ*,p)* $\wedge$ *Diminution(*τ*)* $\wedge$ *Digeste(p).*

Soit $a = (\varphi_0, \ldots, \varphi_k)$ un argument. Nous notons $Supp(a) = \varphi_0$ le support de a et $Conc(a) = \varphi_k$ la conclusion de a. Etant donnée la base de connaissances $\mathcal{K}_{agg}$ nous notons $\mathcal{A}$ l'ensemble des arguments construits à partir des faits de $\mathcal{K}_{agg}$. Nous définissons alors la *relation d'attaque Att* comme un sous-ensemble de $\mathcal{A} \times \mathcal{A}$. Soient a et b deux arguments. On dit que l'argument a attaque l'argument b, c'est-à-dire $(a, b) \in Att$, si et seulement si $\exists \varphi \in Supp(b)$ tel que l'ensemble des faits $\{Conc(a), \varphi\}$ est $\mathcal{R}$-inconsistant.

Exemple 5.12 *Soit a l'argument de l'exemple 5.11 et b l'argument suivant :* $b = (\{F_4\}, F_4, R_3(F_4))$ *avec* $R_3(F_4) =$ *Pain(p)* $\wedge$ *TauxExtraction(*τ*,p)* $\wedge$ *Augmentation(*τ*)* $\wedge$ *RicheOligoéléments(p).*

L'argument a attaque l'argument b car :
$\{Conc(a), \bigwedge Supp(b)\} = \{R_1(F_2), F_4\}$ *viole la contrainte négative N.*

Dans ce cadre, un *système d'argumentation* est un couple $(\mathcal{A}, Att)$ où $\mathcal{A}$ est un ensemble d'arguments sur $\mathcal{K}_{agg}$ et Att est la relation d'attaque définie sur cet ensemble.

Exemple 5.13 *Poursuivons l'exemple précédent. Nous considérerons par la suite le système d'argumentation* $(\mathcal{A}, Att)$ *où* $\mathcal{A}$ *est constitué des arguments suivants :*
$a = (\{F_2\}, F_2, R_1(F_2))$ *avec* $R_1(F_2) =$ *Pain(p)* $\wedge$ *TauxExtraction(*τ*,p)* $\wedge$ *Diminution(*τ*)* $\wedge$ *Digeste(p)*
$b = (\{F_4\}, F_4, R_3(F_4))$ *avec* $R_3(F_4) =$ *Pain(p)* $\wedge$ *TauxExtraction(*τ*,p)* $\wedge$ *Augmentation(*τ*)* $\wedge$ *RicheOligoéléments(p)*
$c = (\{F_2\}, F_2, R_4(F_2))$ *avec* $R_4(F_2) =$ *Pain(p)* $\wedge$ *TauxExtraction(*τ*,p)* $\wedge$ *Diminution(*τ*)* $\wedge$ *SansPesticides(p)*
$d = (\{F_3\}, F_3, R_2(F_3))$ *avec* $R_2(F_3) =$ *Pain(p)* $\wedge$ *AdditionSel(s,p)* $\wedge$ *Diminution(s)* $\wedge$ *PeuSalé(p)*
et $Att = \{(a, b), (b, a), (b, c), (c, b)\}$.

Remarque 5.1 *Notons que dans l'exemple ci-dessus la relation d'attaque est symétrique, ce qui n'est pas nécessairement le cas. Considérons les faits suivants :*
$F'_1 =$ *Pain(bleuette)* $\wedge$ *SansContaminants(bleuette)*
$F'_2 = \exists$ *f (TeneurFibres(f,bleuette)* $\wedge$ *Elevé(f)),*
les règles suivantes :
$R'_1 = \forall$ *x (SansContaminants(x)* $\rightarrow$ *SansPesticides(x)* $\wedge$ *SansMycotoxines(x))*
$R'_2 = \forall$ *x,y (TeneurFibres(y,x)* $\wedge$ *Elevé(y))* $\rightarrow$ *ContientContaminants(x))*
et la contrainte négative suivante :
$N' = \exists$ *x (ContientContaminants(x)* $\wedge$ *SansContaminants(x)).*
Soient a' *et* b' *les arguments construits comme suit :*
$a' = (\{F'_1\}, F'_1, R'_1(F'_1))$
$b' = (\{F'_2\}, F'_2, R'_2(F'_2))$ *avec* $R'_2(F'_2) = \exists$ *f (TeneurFibres(f,bleuette)* $\wedge$ *Elevé(f)* $\wedge$ *ContientContaminants(bleuette)).*
L'argument b' *attaque l'argument* a' *car :*
$\{Conc(b'), \bigwedge Supp(a')\} = \{R'_2(F'_2), F'_1\}$ *viole la contrainte négative* N'.
On obtient une relation d'attaque non symétrique : $Att' = \{(b', a')\}$.

Rappelons enfin les principes sur lesquels s'appuie la notion de "cohérence" en argumentation et qui vont nous permettre, au moyen des extensions, de calculer les sous-ensembles d'arguments maximaux consistants dans la dernière étape de notre démarche.

Soit $(\mathcal{A}, Att)$ un système d'argumentation, soient $\mathcal{B} \subseteq \mathcal{A}$ et $a \in \mathcal{A}$. On dit que $\mathcal{B}$ est *sans conflit* ssi il n'existe pas d'arguments $a, b \in \mathcal{B}$ tels que $(a, b) \in Att$. On dit que le sous-ensemble $\mathcal{B}$ *défend* l'argument a ssi pour tout

argument $b \in \mathcal{A}$, si $(b, a) \in Att$ alors il existe $c \in \mathcal{B}$ tel que $(c, b) \in Att$. Le sous-ensemble $\mathcal{B}$ est dit *admissible* ssi il est sans conflit et défend tous ses arguments.

Différentes sémantiques ont été définies pour la notion d'extension. $\mathcal{B}$ est une *extension complète* ssi $\mathcal{B}$ défend tous ses arguments et contient tous les arguments qu'il défend. $\mathcal{B}$ est une *extension préférée* ssi c'est un ensemble admissible maximal (au sens de l'inclusion ensembliste). $\mathcal{B}$ est une *extension stable* s'il est sans conflit et pour tout $a \in \mathcal{A} \setminus \mathcal{B}$, il existe un argument $b \in \mathcal{B}$ tel que $(b, a) \in Att$. $\mathcal{B}$ est une *extension semi-stable* ssi $\mathcal{B}$ est une *extension complète* et que l'union entre l'ensemble $\mathcal{B}$ et l'ensemble de tous les arguments attaqués par $\mathcal{B}$ est maximal (pour l'inclusion ensembliste). L'ensemble des extensions suivant une sémantique x donnée est noté $Ext_x(\mathcal{A}, Att)$.

Comme antérieurement prouvé en logique propositionnelle (et facilement étendu au sous-ensemble de la logique du premier ordre étudié dans ce chapitre), les sous-ensembles de connaissances consistants maximaux correspondent aux extensions (confondues) des sémantiques stable, semi-stable et préférée dans un système d'argumentation (Vesic, 2012).

Exemple 5.14 *Dans le système d'argumentation de l'exemple 5.13, les sémantiques stable, semi-stable et préférée (confondues) ont pour extensions :* $Ext_{stable}(\mathcal{A}, Att) = Ext_{semi-stable}(\mathcal{A}, Att) = Ext_{preferee}(\mathcal{A}, Att) = \{\{a, c, d\}, \{b, d\}\}$.

5.4.2 Choix des points de vue à retenir

Partant de l'ensemble des extensions $Ext_x(\mathcal{A}, Att)$, la démarche d'aide à la décision est la suivante :

1. Pour chaque extension $\varepsilon \in Ext_x(\mathcal{A}, Att)$:
   - considérer les faits apparaissant dans les arguments de ε ;
   - identifier les bases de connaissances $\mathcal{K}_i$ dont ces faits sont issus ;
   - obtenir ainsi l'ensemble des buts G_i satisfaits ;
   - déduire l'ensemble des points de vue associés, à l'aide de la fonction κ ;
   - présenter aux experts l'ensemble des buts et des points de vue compatibles, correspondant à l'extension considérée.
2. On obtient de cette manière autant d'options qu'il y a d'extensions dans $Ext_x(\mathcal{A}, Att)$. Pour décider de l'option à retenir, plusieurs lignes de

conduite sont alors envisageables et proposées aux experts. Le processus de décision peut notamment :

- suggérer l'option satisfaisant le plus grand nombre de buts ;
- suggérer l'option contribuant à satisfaire le plus grand nombre de points de vue ;
- demander à l'expert du domaine de définir une préférence sur les buts et/ou les points de vue considérés.

Exemple 5.15 *Dans notre exemple la première extension $\{a, c, d\}$ s'appuie sur les faits F_2 et F_3, issus des bases de connaissances $\mathcal{K}_1$, $\mathcal{K}_2$ et $\mathcal{K}_4$ et satisfaisant les buts G_1, G_2 et G_4. Les buts G_1 et G_2 sont associés au point de vue nutritionnel, tandis que le but G_4 est associé au point de vue sanitaire.*

La seconde extension $\{b, d\}$ s'appuie sur les faits F_3 et F_4, issus des bases de connaissances $\mathcal{K}_2$ et $\mathcal{K}_3$ et satisfaisant les buts G_2 et G_3, tous deux associés au point de vue nutritionnel.

La première option (correspondant à l'extension $\{a, c, d\}$) consisterait à réaliser les faits F_2 et F_3 et permettrait de satisfaire le plus grand nombre de buts et de prendre en compte le plus grand nombre de points de vue.

La seconde option (correspondant à l'extension $\{b, d\}$) consisterait à réaliser les faits F_3 et F_4. Elle satisfairait deux buts et prendrait en compte un seul point de vue. Elle peut toutefois être pertinente si le but G_3, qui n'est pas satisfait par la première option, est jugé prioritaire.

5.5 Synthèse et discussion

5.5.1 Schéma global de la démarche

Une synthèse de la démarche proposée est schématisée sur la figure 5.5.1 :

• L'étape 0 désigne la formalisation des buts, des points de vue, et leur correspondance par la fonction κ comme présenté dans la section 5.3.3.

• L'étape 1 correspond au calcul, par dérivation inverse, des bases de connaissances associées à chaque but (section 5.3.5). On considère alors leur union.

• L'étape 2 correspond au calcul des arguments et des extensions tel qu'exposé dans la section 5.4.1.

• Enfin l'étape 3 détermine le choix des points de vue conformément à la section 5.4.2.

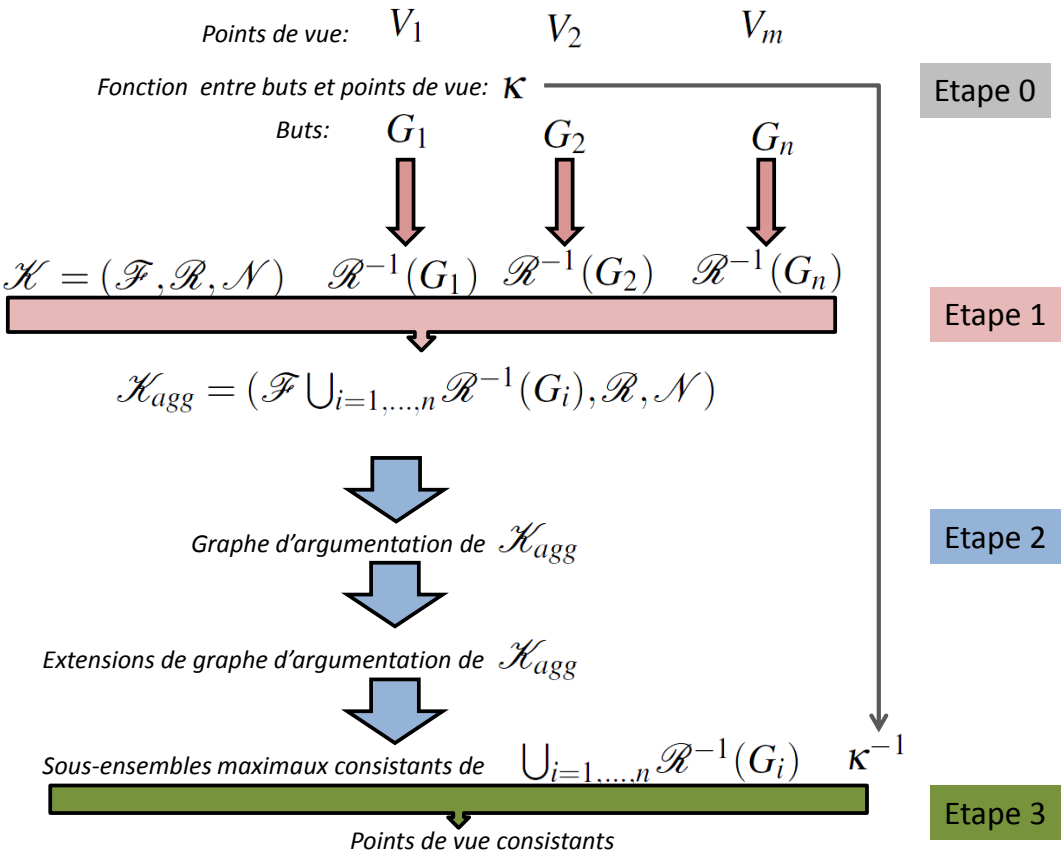

FIGURE 5.4 – Schéma global de la démarche

5.5.2 Autres approches et positionnement

La conciliation des différents points de vue en jeu au sein d'une filière soulève des questions encore peu abordées formellement. Les liens entre argumentation et décision sont peu explorés dans la littérature, la recherche en décision argumentée ayant été amorcée assez récemment avec les travaux de (Amgoud et Prade, 2009). Il existe par ailleurs une communauté internationale active sur différents aspects de l'argumentation (Besnard et Hunter, 2008; Rahwan et Simari, 2009) d'une part, et sur la décision multicritère (Bouyssou et collab., 2009) d'autre part. Cependant ces deux domaines ont essentiellement été abordés séparément jusqu'ici. Or la démarche paraît extrêmement pertinente pour aider à la prise de décision dans les filières agroalimentaires (Thomopoulos et collab., 2009), dans un objectif d'arbitrage.

Le contexte d'application principal de l'argumentation – au sens large, et non au sens d'un cadre formel en informatique, initié par (Dung, 1995) – est historiquement le domaine judiciaire, bien que d'autres applications aient été abordées du point de vue du management ou de l'étude de débats en sciences de gestion (Maguire et Boiney, 1994; Gottsegen, 1998).

Récemment, les premiers travaux en argumentation pour la prise de décision appliqués à l'arbitrage au sein de la filière céréalière ont été réalisés (Bourguet, 2010; Bourguet et collab., 2013). Ils s'appuient sur un modèle de décision argumentée pour proposer des recommandations de consommation destinées à des publics spécifiques (prévention cardiovasculaire, etc.). Le présent chapitre aborde un verrou essentiel pour le pilotage de filière en y associant l'ingénierie inverse et introduit une démarche permettant de proposer des options en termes de procédés de fabrication. Notons qu'en revanche l'argumentation est utilisée ici comme mode de calcul, et non comme moyen d'analyse, des objectifs compatibles dans la filière.

Comme discuté dans les section précédentes, la technique de chaînage arrière en elle-même est utilisée dans d'autres travaux tels que (Baget et Salvat, 2006; Klarman et collab., 2011). L'approche présentée dans ce chapitre se rapproche de (Baget et Salvat, 2006), qui s'intéresse à la réponse à une requête, si l'on considère un but comme une requête et les moyens d'atteindre ce but comme une réponse. Toutefois ici cette réponse n'est pas nécessairement dans la base de connaissances initiale mais calculée par dérivation inverse. La démarche présente aussi des points communs avec (Klarman et collab., 2011), par le calcul d'un ensemble de faits minimal. Toutefois les ensembles de faits

sont obtenus ici par les extensions, qui cherchent au contraire à maximiser l'ensemble des arguments compatibles.

Enfin une série de travaux tels que (Amgoud et Cayrol, 2002; Bench-Capon, 2003; Kaci et van der Torre, 2008; Amgoud et collab., 2000; Bourguet et collab., 2013) ont introduit dans les systèmes d'argumentation la notion de force des arguments, par le biais de préférences ou de valeurs associées aux arguments. En revanche ces travaux ne se situent pas dans une optique d'ingénierie inverse et ne manipulent pas les éléments du formalisme logique utilisé ici. Il serait de ce fait judicieux d'étudier l'introduction de préférences sur les buts, les règles ou les faits. L'expression de préférences pourrait alors participer à la résolution des conflits.

5.6 Conclusion du chapitre

Cette étude de cas représente une application originale et une approche introspective dans le domaine agroalimentaire et l'aide à la décision pour les filières. Il nécessite néanmoins une tâche de modélisation des connaissances très coûteuse, qui ne peut en l'état être automatisée, et qui dépend fortement de la qualité de l'expertise et de l'élicitation (exhaustivité, certitude, etc.). La tendance des outils d'aide à la décision inclut de plus en plus les méthodes d'argumentation comme moyen d'impliquer les parties prenantes dans la tâche de modélisation et le processus décisionnel en favorisant leurs interactions.

Nous avons pris le parti dans ce chapitre de nous focaliser sur la démarche adoptée, plutôt que de présenter l'application de façon exhaustive. Une phase visant à la fois à compléter les connaissances saisies et à évaluer l'intérêt du système est actuellement en cours, sous la forme d'une série d'ateliers permettant les échanges avec et entre les experts de la filière. Elle porte sur un nombre restreint de scénarios préalablement identifiés comme controversés. L'optimisation algorithmique est une perspective importante de ce travail. Un autre élément à prendre en compte dans la prise de décision est l'expression de préférences sur les éléments du formalisme, reflétant l'importance relative des buts, des règles ou des faits du système. Une façon de modéliser cet aspect dans le formalisme reste à étudier.

Comme précédemment indiqué, nous avons considéré un sous-ensemble du langage logique introduit par (Chein et Mugnier, 2009) afin de nous confor-

mer au mieux au cas d'utilisation étudié. Il serait intéressant d'utiliser cette méthodologie dans d'autres scénarios où ce type de règles se présentent naturellement. Une première étape est bien sûr l'identification de tels scénarios. Puis les notions de dérivation inverse et dérivation avant doivent être étendues afin de permettre des dérivations potentiellement infinies (un résultat direct de l'introduction de variables existentielles en conclusion). La notion de minimalité telle qu'abordée par (Klarman et collab., 2011) est alors essentielle.

Sur les aspects touchant l'argumentation, il serait également judicieux d'étudier les conséquences de l'expression de préférences sur les points de vue. Comment de telles préférences se traduiront en préférences entre arguments est encore une question ouverte.

Chapitre 6

Perspectives

Ce chapitre est consacré aux perspectives qui se dégagent de ces travaux. Elles sont décrites ci- dessous suivant trois orientations majeures :

- l'étude d'une approche graphique et collaborative de la décision argumentée ;
- les systèmes d'argumentation pour l'analyse multidimensionnelle d'un système ;
- la fusion de données redondantes.

6.1 Décision argumentée : vers une approche graphique et collaborative

Le chapitre 3 a montré, dans le domaine de la prédiction, l'intérêt d'une approche collaborative faisant intervenir les experts de façon approfondie. Il s'agit d'une approche ouverte à la critique et permettant aux spécialistes du domaine d'"injecter" des connaissances qui leur apparaissent comme des lacunes du modèle nuisant à sa justesse. Ses principaux points forts sont l'interprétabilité des modèles résultants et l'interaction avec les experts qui favorise leur implication et leur confiance dans les modèles développés.

Dans un domaine tel que l'argumentation pour l'analyse d'une situation complexe et l'aide à la décision, l'adhésion des parties prenantes à la démarche est cruciale. D'où l'intérêt d'adopter ce type d'approche dans le domaine de la décision argumentée présentée dans le chapitre 4. Une des spécificités du

modèle présenté dans le chapitre 3, favorisant l'interaction avec les experts, est son aspect graphique. Nous n'avons en revanche pas développé l'aspect graphique dans le chapitre 4.

Il existe à l'heure actuelle deux principaux outils de visualisation d'arguments : Araucaria (http ://araucaria.computing.dundee.ac.uk/) et Carneades (http ://carneades.berlios.de/downloads/). Ces outils facilitent la visualisation des arguments et en particulier de leur structure (prémisses, conclusion, inférence) mais ce ne sont pas des outils de raisonnement. De plus, ils ne sont pas orientés vers l'aide à la décision et ne prennent donc pas en compte les éléments de la décision.

Dans le chapitre 4, nous avons pris le parti d'associer à chaque argument une action et un but. La relation d'attaque entre arguments est alors définie à partir de l'exclusion mutuelle ou de la spécialisation des actions associées, tandis que la relation de préférence entre arguments est définie à partir de la priorité accordée aux buts associés. Ces notions sont également utilisées dans le chapitre 5 (sous les termes : options, buts, inconsistance). L'équipe a ébauché dans (Bourguet, 2010; Fortin et collab., 2011) un formalisme graphique et logique fondé sur les graphes conceptuels pour la représentation d'arguments et des mécanismes de raisonnement impliqués. Plusieurs aspects de l'argumentation y sont pris en compte.

D'une part, une structure interne de l'argument y est définie, prenant en compte les éléments de la décision (raison, but, action) et proposant l'introduction d'une ontologie. Sur cette base, la relation d'attaque s'appuyant sur l'incompatibilité des actions est introduite, permettant de générer un graphe d'attaque.

D'autre part, le calcul de différents types d'extensions est défini par des règles de graphes conceptuels avec "défauts". En effet, l'équipe a travaillé sur une extension du modèle pour permettre de réaliser des raisonnements non-monotones (logique des défauts) (Baget et collab., 2009; Baget et Fortin, 2010). On désigne par là un raisonnement qui permet de remettre en cause une déduction déjà faite lorsqu'arrive une information nouvelle. Il permet typiquement de prendre en compte des exceptions à des règles générales ou du raisonnement à différents niveaux de précision. (Fortin et collab., 2011) a montré que l'utilisation de raisonnements non-monotones fournit un paradigme déclaratif graphique simple et intuitif pour le calcul des ensembles d'arguments maximaux cohérents. L'ajout d'un nouvel argument dans le système remet en question d'autres arguments jusque-là acceptés.

Certains travaux antérieurs, bien que peu nombreux (Moulin et Irandoust, 1999; Moor et collab., 2009), se sont intéressés à l'analyse du discours argumentatif par les graphes conceptuels, sans toutefois y introduire les mécanismes de raisonnement permettant l'analyse de la situation ou l'aide à la décision.

L'approche initiée dans (Bourguet, 2010; Fortin et collab., 2011) peut servir de base pour permettre à l'avenir d'approfondir les aspects graphiques du processus de décision. Une des premières étapes consiste à étudier la représentation d'une base d'arguments comme un réseau mettant facilement en évidence des éléments-clés communs tels que les buts communs à plusieurs parties prenantes, les actions sur lesquelles elles s'accordent et à l'inverse, les éléments de désaccord et les raisons sous-jacentes. En fonction du mode de décision adopté (vote, priorité à certains buts, etc.), les actions associées pourront être mises en évidence graphiquement. La possibilité de calcul de ces résultats par des opérations de graphes pourra être étudiée.

Par ailleurs, les travaux développés jusqu'ici s'appuient principalement sur la synthèse d'arguments présentés dans différentes sources de données (articles, entrevues d'experts, rapports). Plus récemment, ils ont intégré des discussions entre experts, introduisant une certaine interactivité. Toutefois la démarche présentée dans le chapitre 3 a démontré l'utilité d'une collaboration étroite et interactive avec les utilisateurs du modèle pour faire émerger et formaliser des connaissances manquantes pertinentes pour l'analyse et la compréhension du problème. De plus, la prise de décision *in fine* concernant le choix d'options mises en avant par une autorité (gouvernementale en l'occurrence), ne pourra se faire qu'avec l'implication des parties prenantes.

Au sein ou au-delà des éléments de la décision pris en compte dans la structure interne des arguments (raison, but, action), on peut penser qu'il est possible de réaliser avec chaque partie prenante un affinement de la base d'arguments en identifiant et en modélisant des facteurs (explications, interactions, ...) considérés comme significatifs. La démarche étant itérative, les étapes de vérification et validation permettent éventuellement de compléter et d'enrichir le modèle.

Les différents acteurs impliqués ayant des points de vue différents et des visions partielles du problème, une question de recherche consiste à explorer l'intégration des différents modèles dans une vision globale reflétant la pluralité des points de vue des parties prenantes. Divers aspects sont envisageables :

- les éléments apportés par l'acteur correspondent-ils à des éléments du système d'argumentation et de décision ?
- peuvent-ils être modélisés de façon générique, en fonction de la nature de la connaissance apportée (explication, alternative, ...) ?
- y a-t-il une raison pour que ces éléments n'aient pas été identifiés d'emblée (arguments "cachés", ...) ?
- ces éléments ont-ils des liens avec les autres parties prenantes ? Sont-ils suscités par des connaissances apportées par d'autres acteurs ou en suscitent-ils ?
- modifient-ils la sortie du système d'argumentation ?

A plus long terme, dans une situation de débat pour la recherche (idéalement) d'un consensus, on peut imaginer que les acteurs puissent être amenés à changer d'opinion, retirer ou assouplir certains de leurs arguments ou en apporter d'autres. Nous faisons l'hypothèse que la connaissance du modèle global contribuerait à faire émerger une solution globalement acceptable : d'une part, parce qu'elle permettrait de présenter à chaque partie prenante des arguments intéressant son point de vue ; d'autre part, parce que chaque partie prenante serait plus facilement à même d'identifier les changements de position qui permettraient de débloquer la situation. Ces considérations nous mènent donc vers une thématique proche mais différente : la négociation.

6.2 Argumentation et analyse multidimensionnelle

Les systèmes d'argumentation ont probablement un rôle à jouer dans l'analyse de systèmes tels que les filières selon différentes dimensions qualitatives, telles que prises en compte par exemple dans les sciences sociales. L'introduction de contextes dans les systèmes d'argumentation peut servir de base à une telle approche.

Des systèmes d'argumentation avec contextes ont déjà été proposés dans la littérature, permettant de représenter plusieurs dimensions ou points de vue sur un problème. Une série de travaux ont d'abord étendu le cadre de Dung (1995) pour introduire des préférences dans un système d'argumentation. Dans Amgoud et Cayrol (2002), une relation binaire de préférence entre arguments est proposée. Cette relation capture les différences de force entre arguments et peut procéder à une sélection entre arguments conflictuels. Une

proposition analogue est introduite par Bench-Capon (2003), qui associe à chaque argument une valeur (ou plusieurs dans Kaci et van der Torre (2008)) et définit une préférence entre ces valeurs, induisant une relation d'ordre entre arguments.

Dans ces travaux, les préférences sont supposées non contradictoires. Cependant, dans les applications courantes, cela n'est pas toujours le cas. D'où l'introduction de contextes. Dans (Amgoud et collab., 2000), une extension de (Amgoud et Cayrol, 2002) est proposée avec l'idée que l'ensemble $\mathcal{A}$ des arguments peut être muni de plusieurs relations de préférence, chacune d'elle exprimant des préférences entre arguments dans un contexte particulier. Une phase d'agrégation est ensuite nécessaire pour combiner les préférences des différents contextes. D'autres extensions avec contextes, telles que (Bourguet et collab., 2013), ont été proposées afin d'élargir l'expressivité du système d'argumentation.

Nous pensons judicieux de définir un cadre qui ne comporte pas de préférences, afin de n'émettre aucun jugement de valeur concernant l'importance des différentes dimensions étudiées, mais comportant en revanche plusieurs séries de contextes, dans le but d'identifier les contextes les plus significatifs. L'ensemble des arguments serait partitionné par les différents contextes.

A titre d'illustration, une étude préliminaire présentée dans (Thomopoulos et collab., 2013b) s'est intéressée à deux partitions d'un ensemble d'arguments. La première (figure 6.1) est réalisée selon les contextes économique, technique, social et participatif, la seconde (figure 6.2) selon les contextes favorable ou non favorable à la mise en place d'un dispositif d'approvisionnement en circuit court. Le système admet deux extensions préférées. Les arguments appartenant à la première d'entre elles apparaissent en vert dans les deux figures.

Une perspective qui semble pertinente pour l'analyse d'un système serait la définition d'une méthodologie permettant l'identification des contextes les plus significatifs dans la compréhension du système et la réponse à des questions telles que :

- quels sont les contextes les plus polémiques ?
- la polémique est-elle issue d'un contexte présentant des conflits internes ou de la confrontation de plusieurs contextes ?
- quels sont les contextes les plus consensuels ?
- quelle partition permet de mieux discriminer les arguments acceptables ?

Une telle approche pourrait s'appuyer notamment sur l'étude du système

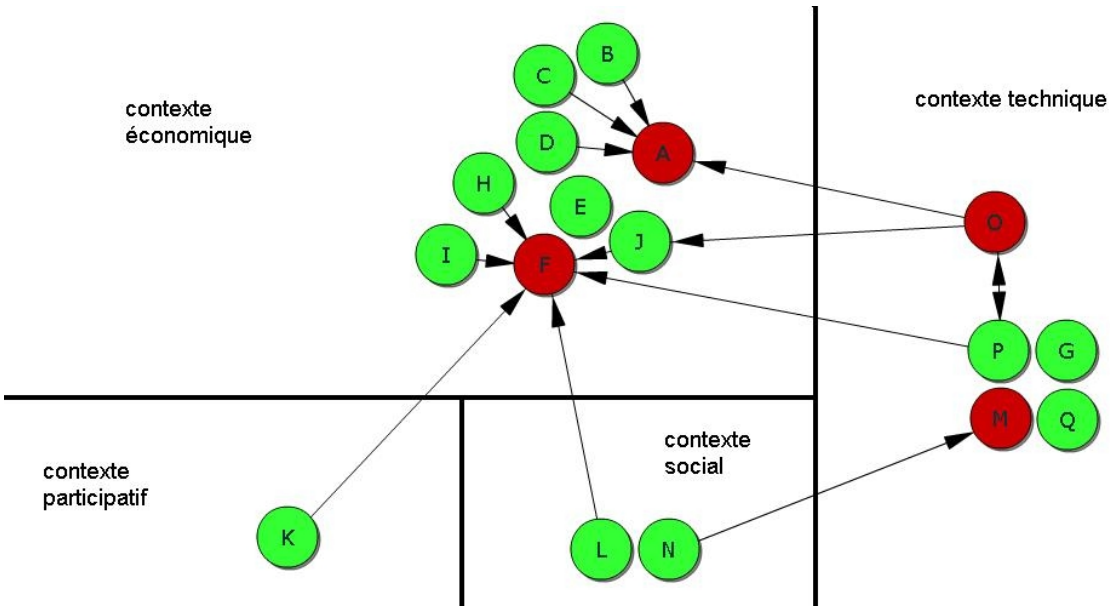

FIGURE 6.1 – Première partition

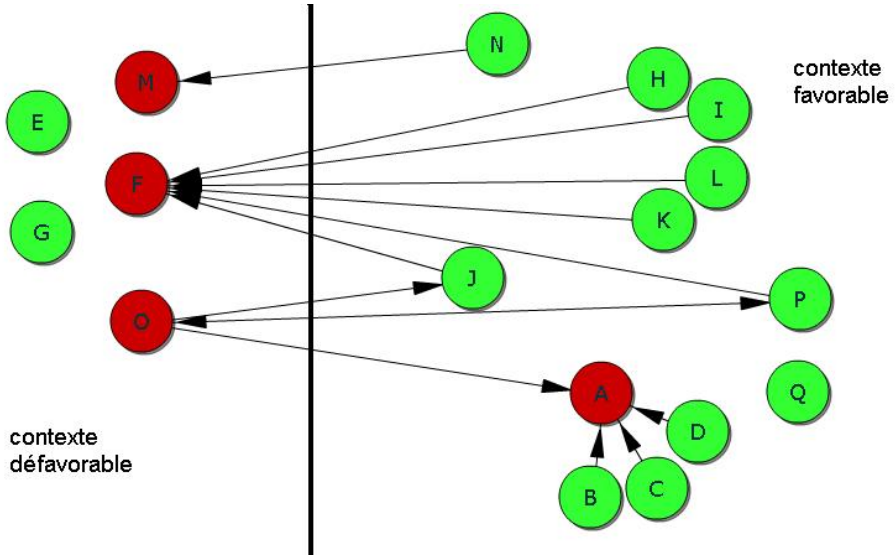

FIGURE 6.2 – Seconde partition

réduit à chacun des contextes et comparé au système global au moyen d'indicateurs (par exemple, nombre d'extensions dans le système global et dans

chaque contexte, nombre d'attaques internes/externes, etc.). Les systèmes d'argumentation pourraient ainsi fournir une méthode d'évaluation qualitative de systèmes tels que les filières.

6.3 Qualité des données et fusion de données redondantes

Cette perspective se situe dans le thème de l'intégration de connaissances. Elle concerne le cas, non présenté dans le chapitre 2, où des données venant de plusieurs sources (bases de données, ...) réfèrent à la même entité du monde réel et sont donc redondantes. Dans de tels cas, il est nécessaire, d'une part, d'identifier ces références multiples, et d'autre part de synthétiser cette information redondante de manière à ce qu'elle soit facilement manipulable.

Nous supposons résolu le problème de détection de redondances (Saïs et collab., 2007), qui consiste à décider, sur la base des informations contenues dans les données (ou références), si deux références distinctes représentent la même entité du monde réel (par exemple le même auteur, le même livre, etc.). Une fois les groupes de références redondantes formés, se présente le problème de fusion des références au sein de chaque groupe en une référence unique. C'est à ce problème que nous nous intéressons ici. Le fait de fusionner les références redondantes a plusieurs intérêts : d'une part il permet de réduire le nombre total de références, ce qui permet un stockage plus facile et une interrogation plus rapide des données, d'autre part la réponse renvoyée suite à une requête d'utilisateur est plus lisible, les doublons ayant été fusionnés.

Les valeurs prises par un même attribut au sein d'un groupe de références réconciliées sont variables (par exemple, un tableau est appelé tantôt *Mona Lisa*, tantôt *Joconde*), contiennent des erreurs (pouvant être introduites durant le processus d'acquisition des données) et peuvent être incomplètes (attributs non renseignés). La plupart des systèmes de fusion proposés dans la littérature demandent une intervention de l'utilisateur et renvoient des références fusionnées où chaque attribut est doté d'une valeur unique (Papakonstantinou et collab., 1996; Subrahmanian et collab., 1995). Dans (Saïs et Thomopoulos, 2008b,a; Destercke et collab., 2009; Saïs et collab., 2010), une approche possibiliste a été développée pour prendre en compte cette incertitude lors de la construction de la référence fusionnée. Elle s'appuie sur différents critères (fiabilité des sources, fréquence d'occurrence, similarité

syntaxique, ...).

Cette thématique sera développée dans le cadre du projet ANR "Qualinca" qui concerne la qualité et l'interopérabilité de grands catalogues documentaires. Elle permettra d'explorer plusieurs axes :

L'expérimentation sur de grandes bases documentaires offre la possibilité, d'une part, d'évaluer la méthode sur de grands volumes de données, d'autre part d'étudier la prise en compte d'une ontologie riche du domaine d'application.

Elle permet également d'établir une comparaison expérimentale de la pertinence des différents critères utilisés et d'en envisager d'autres, comme la prise en compte du caractère plus ou moins informatif des valeurs considérées, à l'aide notamment de la relation de spécialisation, ou l'apport d'informations externes issues du web.

Enfin nous envisageons d'étendre la méthode proposée à la prise en compte de données multivaluées. Par exemple, l'attribut "métier" (et bien d'autres) peut prendre plusieurs valeurs. Dans le cas de telles données multivaluées, plusieurs valeurs parmi celles figurant dans les données peuvent être les bonnes, et la référence fusionnée ne fournit plus un ensemble de valeurs exclusives comme supposé dans le modèle jusqu'ici proposé.

Bibliographie

Adomavicius, G. et A. Tuzhilin. 2001, «Expert-driven validation of rule-based user models in personalization applications», *Data Mining and Knowledge Discovery*, vol. 5, n° 1-2, p. 33–58.

Amgoud, L. et C. Cayrol. 2002, «A reasoning model based on the production of acceptable arguments», *Annals of Mathematics and Artificial Intelligence*, vol. 34, p. 197–216.

Amgoud, L., S. Parsons et L. Perrussel. 2000, «An argumentation framework based on contextual preferences», dans *Proceedings of the International Conference on Formal and Applied and Practical Reasoning*, p. 59–67.

Amgoud, L. et H. Prade. 2009, «Using arguments for making and explaining decisions», *Artificial Intelligence*, vol. 173, n° 3-4, p. 413–436.

AQUANUP. 2009, `http://www.inra.fr/inra_cepia/vous_recherchez/des_projets/france/aquanup`.

Baget, J.-F., M. Croitoru, J. Fortin et R. Thomopoulos. 2009, «Default conceptual graph rules : preliminary results for an agronomy application», *Proc. of ICCS'09, Lecture Notes in Artificial Intelligence*, vol. 5662, p. 86–99.

Baget, J.-F. et J. Fortin. 2010, «Default conceptual graph rules, atomic negation and tic-tac-toe», *Proc. of ICCS'10, Lecture Notes in Artificial Intelligence*, vol. 6208, p. 42–55.

Baget, J.-F. et E. Salvat. 2006, «Rules dependencies in backward chaining of conceptual graphs rules», dans *Conceptual Structures : Inspiration and Application, 14th International Conference on Conceptual Structures*, *LNCS*, vol. 4068, Springer, p. 102–116.

Baroni, P., M. Giacomin et G. Guida. 2005, «Scc-recursiveness : a general schema for argumentation semantics», *Artificial Intelligence Journal*, vol. 168 (1-2), p. 162–210.

Ben-David, A. et L. Sterling. 2006, «Generating rules from examples of human multiattribute decision making should be simple», *Expert Syst. Appl.*, vol. 31, n° 2, p. 390–396.

Bench-Capon, T. J. M. 2003, «Persuasion in practical argument using value-based argumentation frameworks», *Journal of Logic and Computation*, vol. 13, n° 3, p. 429–448.

Besnard, P. et A. Hunter. 2008, *Elements of Argumentation*, The MIT Press.

Bonet, B. et H. Geffner. 1996, «Arguing for decisions : A qualitative model of decision making», dans *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI'96)*, p. 98–105.

Bos, C., B. Botella et P. Vanheeghe. 1997, «Modeling and Simulating Human Behaviors with Conceptual Graphs», dans *Proc. of ICCS'97*, *LNAI*, vol. 1257, Springer, p. 275–289.

Bourguet, J.-R. 2010, *Contribution aux méthodes d'argumentation pour la prise de décision. Application à l'arbitrage au sein de la filière céréalière*, Thèse de doctorat, Université Montpellier II, Montpellier, France.

Bourguet, J.-R., L. Amgoud et R. Thomopoulos. 2009, «Contribution aux comparaisons formelles des modèles de préférences en argumentation», dans *Journées Francophones des Modèles formels de l'intéraction (MFI)*, p. 81 – 92.

Bourguet, J.-R., L. Amgoud et R. Thomopoulos. 2010, «Towards a unified model of preference-based argumentation», dans *International Symposium of Foundations of Information and Knowledge Systems (FoIKS)*, p. 326–344.

Bourguet, J.-R., R. Thomopoulos, M.-L. Mugnier et J. Abécassis. 2013, «An artificial intelligence-based approach to deal with argumentation applied to food quality in a public health policy», *Expert Systems With Applications*, vol. 40, n° 11, p. 4539–4546.

Bourre, J.-M., A. Bégat, M.-C. Leroux, V. Mousques-Cami, N. Pérandel et F. Souply. 2008, «Valeur nutritionnelle (macro et micro-nutriments) de farines et pains français», *Médecine et Nutrition*, vol. 44, n° 2, p. 49–76.

Bouyssou, D., D. Dubois, M. Pirlot et H. Prade. 2009, *Decision-making process – Concepts and Methods*, Wiley.

CADINNO. 2008, «Information, choix, consommateurs responsables : des leviers pour un développement durable ?», `http://www.melissa.ens-cachan.fr/IMG/pdf/Colloque_CadInno_FR.pdf`.

Caragea, D., J. Zhang, J. Bao, J. Pathak et V. Honavar. 2005, «Algorithms and software for collaborative discovery from autonomous, semantically heterogeneous, distributed information sources», dans *ALT*, *Lecture Notes in Computer Science*, vol. 3734, édité par S. Jain, H.-U. Simon et E. Tomita, Springer, ISBN 3-540-29242-X, p. 13–44.

Chein, M. et M.-L. Mugnier. 2009, *Graph-based Knowledge Representation and Reasoning, Computational Foundations of Conceptual Graphs*, Advanced Information and Knowledge Processing Series, Springer London.

Clocksin, W. F. et C. Mellish. 1984, *Programming in Prolog, 2nd Edition*, Springer.

Dean, M., R. Sheperd, A. Arvola, P. Lampila, L. Lahteenmaki, M. Vassalo, A. Saba, E. Claupein et M. Winkelmann. 2007, «Report on consumer expectations of health benefits of modified cereal products», cahier de recherche, University of Surrey, UK.

Destercke, S., F. Saïs et R. Thomopoulos. 2009, «Fusion évidentielle de références et interrogation flexible», dans *Rencontres francophones sur la logique floue et ses applications (LFA'09)*, p. 15–22.

DINABIO. 2008, «Proceedings of dinabio développement et innovation en agriculture biologique», `http://www.inra.fr/ciag/revue_innovations_agronomiques/volume_4_janvier_2009`.

Dubuisson-Quellier, S. 2006, «De la routine à la délibération. les arbitrages des consommateurs en situation d'achat», *Réseaux*, vol. 135/136, p. 253–284.

Dung, P. M. 1995, «On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games», *Artificial Intelligence Journal*, vol. 77, p. 321–357.

Euzenat, J., T. Le Bach, J. Barrasa, P. Bouquet, J. De Bo, R. Dieng-Kuntz, M. Ehrig, M. Hauswirth, M. Jarrar, R. Lara, D. Maynard, A. Napoli, G. Stamou, H. Stuckenschmidt, P. Shvaiko, S. Tessaris, S. Van Acker et I. Zaihrayeu. 2004, «State of the art on ontology alignment», deliverable 2.2.3, Knowledge web NoE.

FCN. 2009, «Fibres, céréales et nutrition», `http://www.inra.fr/content/view/full/24670029`.

Figueira, J., S. Greco et M. Ehrgott. 2005, *Multiple Criteria Decision Analysis : State of the Art Surveys*, Springer Verlag.

Folch, H., B. Habert, M. Jardino, N. Pernelle, M.-C. Rousset et A. Termier. 2004, «Highlighting latent structure in documents», dans *International Conference on Language Resources and Evaluation (LREC)*, vol. 4, p. 1131,1334.

Fortin, J., R. Thomopoulos, J.-R. Bourguet et M.-L. Mugnier. 2011, «Supporting argumentation systems by graph representation and computation», *Proc. of GKR-IJCAI'11, Lecture Notes in Artificial Intelligence*, vol. 7205, p. 119–136.

Fox, J. et S. Das. 2000, *Safe and Sound. Artificial Intelligence in Hazardous Applications*, AAAI Press, The MIT Press.

Gaines, B. R. et M. L. G. Shaw. 1993, «Eliciting knowledge and transferring it effectively to a knowledge-based system», *IEEE Transactions on Knowledge and Data Engineering*, vol. 5, p. 4–14.

Genest, D. 2000, *Extension du modèle des graphes conceptuels pour la recherche d'informations*, thèse de doctorat, Université Montpellier II.

Ginon, E., Y. Lohérac, C. Martin, P. Combris et S. Issanchou. 2009, «Effect of fibre information on consumer willingness to pay for french baguettes», *Food Quality and Preference*, vol. 20, p. 343–352.

Gottsegen, J. 1998, «Using argumentation analysis to assess stakeholder interests in planning debates», *Computers, Environment and Urban Systems*, vol. 22, n° 4, p. 365 – 379.

Guillaume, S. et B. Charnomordic. 2011, «Learning interpretable fuzzy inference systems with fispro», *Information Sciences*, doi :doi:10.1016/j.ins.2011.03.025, p. In Press,.

Haemmerlé, O., P. Buche et R. Thomopoulos. 2006, «The MIEL system : uniform interrogation of structured and weakly structured imprecise data», *Journal of Intelligent Information Systems*.

Haemmerlé, O. et B. Carbonneill. 1996, «Interfacing a relational database using conceptual graphs», dans *DEXA '96 : Proceedings of the 7th International Workshop on Database and Expert Systems Applications*, IEEE Computer Society, Washington, DC, USA, ISBN 0-8186-7662-0, p. 499.

HEALTHGRAIN. 2009, `http://www.healthgrain.org`.

Johnson, I., J. Abécassis, B. Charnomordic, S. Destercke et R. Thomopoulos. 2010, «Making ontology-based knowledge and decision trees interact : an approach to enrich knowledge and increase expert confidence in data-driven models», *Proc. of KSEM'10, Lecture Notes in Artificial Intelligence*, vol. 6291, p. 304–316.

Kaci, S. et L. van der Torre. 2008, «Preference-based argumentation : Arguments supporting multiple values», *International Journal of Approximate Reasoning*, vol. 48, n° 3, p. 730–751.

Kietz, J.-U., A. Maedche et R. Volz. 2000, «A method for semi-automatic ontology acquisition from a corporate intranet», dans *Proceedings of EKAW-2000 Workshop "Ontologies and Text", Juan-Les-Pins, France, October 2000*, n° 1937 dans Springer Lecture Notes in Artificial Intelligence (LNAI). URL `citeseer.ist.psu.edu/kietz00method.html`.

Klarman, S., U. Endriss et S. Schlobach. 2011, «Abox abduction in the description logic *alc*», *J. Autom. Reasoning*, vol. 46, n° 1, p. 43–80.

Kolaitis, P. G. et M. Y. Vardi. 1998, «Conjunctive-Query Containment and Constraint Satisfaction», dans *Proceedings of PODS'98*.

Kraus, S., K. Sycara et A. Evenchik. 1998, «Reaching agreements through argumentation : a logical model and implementation», vol. 104, p. 1–69.

Layat, T. 2011, «Place du pain dans l'équilibre alimentaire», *Pratiques en nutrition*, vol. 7, n° 26, p. 45–50.

Ling, T., B. H. Kang, D. P. Johns, J. Walls et I. Bindoff. 2008, «Expert-driven knowledge discovery», dans *Proceedings of the fifth international conference on information technology : new generations*, édité par S. Latifi, p. 174–178.

Maguire, L. A. et L. G. Boiney. 1994, «Resolving environmental disputes : a framework incorporating decision analysis and dispute resolution techniques», *Journal of Environmental Management*, vol. 42, n° 1, p. 31 – 48.

Maillot, N. et M. Thonnat. 2008, «Ontology based complex object recognition», *Image and Vision Computing*, vol. 26, p. 102–113.

Mansingh, G., K.-M. Osei-Bryson et H. Reichgelt. 2011, «Using ontologies to facilitate post-processing of association rules by domain experts», *Information Sciences*, vol. 181, n° 3, p. 419 – 434, ISSN 0020-0255.

Miller, G. A. 1956, «The magical number seven, plus or minus two : Some limits on our capacity for processing information», *Psychological Review*, vol. 63, p. 81–97.

Moor, A. D., J. Park et M. Croitoru. 2009, «Argumentation map generation with conceptual graphs : the case for essence», dans *CS-TIW at International Conference on Conceptual Structures 2009*, p. 58–69.

Moulin, B. et H. Irandoust. 1999, «Extending the conceptual graph approach to represent evaluative attitudes in discourse», dans *International Conference on Conceptual Structures*, p. 140–153.

Muggleton, S. et L. D. Raedt. 1994, «Inductive logic programming : Theory and methods», *Journal of Logic Programming*, vol. 19/20, p. 629–679. URL `citeseer.ist.psu.edu/muggleton94inductive.html`.

Mugnier, M.-L. 2000, «Knowledge Representation and Reasoning based on Graph Homomorphism», dans *Proc. ICCS'00, LNAI*, vol. 1867, Springer, p. 172–192.

Papakonstantinou, Y., S. Abiteboul et H. Garcia-Molina. 1996, «Object fusion in mediator systems», dans *VLDB*, San Francisco, CA, USA, ISBN 1-55860-382-4, p. 413–424.

Parekh, V. et J.-P. J. Gwo. 2004, «Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies», dans *International Conference of Information and Knowledge Engineering*, The International MultiConference in Computer Science and Computer Engineering, Las Vegas, NV.

Pernelle, N., M.-C. Rousset et V. Ventos. 2001, «Automatic construction and refinement of a class hierarchy over multi-valued data», dans *Principles and Practice of Knowledge Discovery in Databases (PKDD)*, p. 386–398.

PNNS (documents statutaires). 2010, `http://www.sante.gouv.fr/htm/pointsur/nutrition/pol_nutri4.htm`.

PNNS (site web). 2010, `http://www.mangerbouger.fr/menu-secondaire/pnns/le-pnns`.

Popescu, M. et D. Xu. 2009, *Data Mining in Biomedicine Using Ontologies*, 1re éd., Artech House, Inc., Norwood, MA, USA, ISBN 1596933704, 9781596933705.

Quinlan, J. 1986, «Induction of decision trees», *Machine learning*, vol. 1, no 1, p. 81–106.

Quinlan, J. 1993, *C4. 5 : programs for machine learning*, Morgan Kaufmann.

R Development Core Team. 2009, *R : A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`, ISBN 3-900051-07-0.

Rahwan, I. et G. Simari. 2009, *Argumentation in Artificial Intelligence*, Springer.

Saïs, F., N. Pernelle et M.-C. Rousset. 2007, «L2r : A logical method for reference reconciliation», dans *AAAI*, p. 329–334.

Saïs, F. et R. Thomopoulos. 2008a, «Reference fusion and flexible querying», dans *ODBASE-OTM Conferences*, p. 1541–1549.

Saïs, F. et R. Thomopoulos. 2008b, «Une méthode flexible de fusion de références», dans *4èmes Journées francophones Entrepôts de Données et Analyse en ligne (EDA'08)*, p. 77–84.

Saïs, F., R. Thomopoulos et S. Destercke. 2010, «Ontology-driven possibilistic reference fusion», dans *ODBASE-OTM Conferences*, p. 1079–1096.

Salvat, E. 1998, «Theorem proving using graph operations in the conceptual graphs formalism», dans *Proc. of ECAI'98*, p. 356–360.

Sendall, S. et W. Kozaczynski. 2003, «Model transformation : The heart and soul of model-driven software development», *IEEE Software*, vol. 20, n° 5, doi :http://doi.ieeecomputersociety.org/10.1109/MS.2003.1231150, p. 42–45, ISSN 0740-7459.

Slavin, J. et H. Green. 2007, «Diatery fibre and satiety», *British Nutrition Foundation*, vol. 32(1), p. 32–42.

Sowa, J. F. 1976, «Conceptual Graphs», *IBM Journal of Research and Development*.

Sowa, J. F. 1984, *Conceptual Structures : Information Processing in Mind and Machine*, Addison-Wesley.

Stumme, G., A. Hotho et B. Berendt. 2006, «Semantic web mining : State of the art and future directions», *J. of Web Semantics*, vol. 4, p. 124–143.

Subrahmanian, V., S. Adali, A. Brink, R. Emery, J. L. Lu, A. Rajput, T. J. Rogers, R. Ross et C. Ward. 1995, «Hermes : A heterogeneous reasoning and mediator system», .

Sycara, K. 1990, «Persuasive argumentation in negotiation», *Theory and Decision*, vol. 28, p. 203–242.

Termier, A., M.-C. Rousset et M. Sebag. 2002, « TreeFinder : a First Step towards XML Data Mining », dans *International Conference on Data Mining ICDM02*.

Thomopoulos, R., J. Baget et O. Haemmerlé. 2007, «Conceptual graphs as cooperative formalism to build and validate a domain expertise», *Proc. of ICCS'07, Lecture Notes in Computer Science*, vol. 4604, p. 112–125.

Thomopoulos, R., J. Baget et O. Haemmerlé. 2008, «Coopération de connaissances hétérogènes pour la construction et la validation de l'expertise d'un domaine», *Revue des Nouvelles Technologies de l'Information*, vol. E-12, p. 167–186.

Thomopoulos, R., B. Charnomordic, B. Cuq et J. Abécassis. 2009, «Artificial intelligence-based decision support system to manage quality of durum wheat products», *Quality Assurance and Safety of Crops & Foods*, vol. 1, n° 3, p. 179–190.

Thomopoulos, R. et M. Croitoru. 2013, «Aide à la décision en ingénierie inverse : une approche s'appuyant sur l'argumentation», *Revue d'intelligence artificielle*, vol. 27, n° 4-5, p. 493–513.

Thomopoulos, R., S. Destercke, B. Charnomordic, I. Johnson et J. Abécassis. 2013a, «An iterative approach to build relevant ontology-aware data-driven models», *Information Sciences*, vol. 221, p. 452–472.

Thomopoulos, R., D. Paturel et S. Quéré. 2013b, «Un système d'argumentation avec contextes et indicateurs pour analyser un dispositif d'approvisionnement en circuit court», cahier de recherche, INRA.

Tilley, T. A., R. J. Cole, P. Becker et P. W. Eklund. 2005, *A Survey of Formal Concept Analysis Support for Software Engineering Activities*, LNAI3626, Springer-Verlag.

Van Melle, W., A. Scott, J. Bennett et M. Peairs. 1981, *The EMYCIN manual*, Stanford University Press.

Vesic, S. 2012, «Maxi-consistent operators in argumentation.», dans *ECAI*, *Frontiers in Artificial Intelligence and Applications*, vol. 242, édité par L. D. Raedt, C. Bessière, D. Dubois, P. Doherty, P. Frasconi, F. Heintz et P. J. F. Lucas, IOS Press, ISBN 978-1-61499-097-0, p. 810–815.

Vialette, M., A. Pinon, B. Leporq, C. Dervin et J.-M. Membré. 2005, «Meta-analysis of food safety information based on a combination of a relational database and a predictive modeling tool», *Risk Analysis*, vol. 25, n° 1, p. 75–83.

Young, L. 2007, «Application of Baking Knowledge in Software Systems», dans *Technology of Breadmaking - 2nd edition*, Springer, US, p. 207–222.

Zhang, J., D.-K. Kang, A. Silvescu et V. Honavar. 2006, «Learning accurate and concise naïve bayes classifiers from attribute value taxonomies and data», *Knowledge and Information Systems*, vol. 9, n^{o} 2, p. 157–179.

Zhang, J., A. Silvescu et V. Honavar. 2002, «Ontology-driven induction of decision trees at multiple levels of abstraction», *Lecture Notes in Computer Science*, p. 316–323.